Creating, curating, and evaluating a mitogenomic reference database to improve regional species identification using environmental DNA

Emily Dziedzic[1*], Brian Sidlauskas[1], Richard Cronn[2], James Anthony[3], Trevan Cornwell[3], Thomas A. Friesen[3], Peter Konstantinidis[1], Brooke E. Penaluna[2], Staci Stein[3], Taal Levi[1]

[1] Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, OR 97331, USA

[2] Pacific Northwest Research Station, US Department of Agriculture Forest Service, Corvallis, OR, 97311, USA

[3] Oregon Department of Fish and Wildlife, Corvallis Research Laboratory, Corvallis, OR 97333, USA

*Corresponding Author:

Emily Dziedzic

emily.dziedzic@oregonstate.edu

Abstract

Species detection using eDNA is revolutionizing global capacity to monitor biodiversity. However, the lack of regional, vouchered, genomic sequence information—especially sequence information that includes intraspecific variation—creates a bottleneck for management agencies wanting to harness the complete power of eDNA to monitor taxa and implement eDNA analyses. eDNA studies depend upon regional databases of mitogenomic sequence information to evaluate the effectiveness of such data to detect and identify taxa. We created the Oregon Biodiversity Genome Project to create a database of complete, nearly error-free mitogenomic sequences for all of Oregon's fishes. We have successfully assembled the complete mitogenomes of 313 specimens of freshwater, anadromous, and estuarine fishes representing 24 families, 55 genera, and 128 species and lineages. Comparative analyses of these sequences illustrate that many regions of the mitogenome are taxonomically informative, that the short (~150 bp) mitochondrial "barcode" regions typically used for eDNA assays do not consistently diagnose for species, and that complete single or multiple genes of the mitogenome are preferable for identifying Oregon's fishes. This project provides a blueprint for other researchers to follow as they build regional databases, illustrates the taxonomic value and limits of complete mitogenomic sequences, and offers clues as to how current eDNA assays and environmental genomics methods of the future can best leverage this information.

Introduction

The use of ambient genetic material—environmental DNA (eDNA)—to detect and identify metazoans in soil, air, marine environments, and freshwater habitats is transforming how we monitor biodiversity (Andersen et al., 2012; Clare et al., 2022; Deiner et al., 2016; Drummond et al., 2015; Hauck et al., 2019; Lim et al., 2016; Lynggaard et al., 2022; Port et al., 2016; Valentini et al., 2016; Yamamoto et al., 2017). eDNA detection methods depend on comprehensive reference databases of sequence information for target and nontarget species in the clade of interest. The oft-cited lack of comprehensive, reliably vouchered sequence information for many

species (Bohmann et al., 2014; Collins et al., 2013; Cordier et al., 2021; Porter & Hajibabaei, 2018; Schnell et al., 2010) exposes the need to build these reference databases using standardized sample collection, data and specimen curation, and data-sharing protocols (Goldberg et al., 2016). This contribution documents and details the process of creating such a database for Oregon's freshwater fishes and provides a roadmap for others to follow in similar endeavors.

Molecular taxonomists have recommended microgenomic methods (e.g., metabarcoding, barcoding, and single-species detection) for decades to work around the limitations of morphology-based identification (Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, et al., 2003), but molecular species detection methods do have some drawbacks. By design, they rely on diagnostic sequence information from prototypical specimens to ensure correct identification of genetic material found in environmental samples, and that sequence information is not always available. In addition, the gene- and taxon-specific primers needed to amplify barcode regions introduce a key source of error and bias by design in PCR amplification (Yang et al., 2021) because they select certain DNA sequences over others (Fig 1a) (Deiner, Bik, et al., 2017). This primer bias allows researchers to sort metazoan targets from the microorganismal background but can also lead to unwanted loss of information when targeted populations and species amplify variably. Even minor binding biases among target sequences can affect PCR amplification substantially (Nichols et al., 2018; Piñol et al., 2015; Stadhouders et al., 2010), preventing reliable measurements of species presence and/or relative abundance (Yang et al., 2021). Additionally, if species have not diverged at the locus targeted by a primer set, the assay will neither diagnose those taxa nor properly assess their presence or abundance (Fig 1a). Incomplete mitogenomic sequence information prevents *in silico* verification that primers will bind to species' DNA or that the captured region will correctly diagnose species. In addition, missing sequence data and improper taxonomic assignments hinder accurate species identification when querying eDNA metabarcoding results.

Comprehensive databases of error-free, taxonomically verified, full mitogenomic data can solve issues related to unreliable genetic data and greatly improve the accuracy of novel environmental genomics methods that involve sequencing all the DNA in an environmental sample. Such approaches are known as "shotgun sequencing", "ecogenomics", or "community genomics" (Béjà, 2004; Bragg & Tyson, 2014; Taberlet et al., 2012). Researchers focusing on animals typically target areas within the mitochondrial genome ("mitogenome") for eDNA applications because mitochondrial DNA frequently diagnoses taxa accurately (Hebert, Cywinska, et al., 2003), resists environmental degradation (Foran, 2006), and is more easily recovered from degraded samples than lower copy nuclear DNA (Hartmann et al., 2011). Once isolated and sequenced, whole mitogenomes can be used to assign taxonomic identifications in mitochondrial metagenomics (Crampton-Platt et al., 2016), multilocus metabarcoding (Arulandhu et al., 2017; Curd et al., 2019)—where multiple barcode markers are used to identify taxa in a sample—and "ultra-barcoding" (Kane et al., 2012) also known as "super-barcoding" (Li et al., 2015) where much longer barcodes or entire organelles are targeted. Mitogenomic approaches like these can help overcome key challenges with metabarcoding such as primer mismatches, which lead to taxonomic dropout (Cristescu & Hebert, 2018; Piñol et al., 2015; Sipos et al., 2010), reduced quantitative information (Bru et al., 2008; Wu et al., 2009), and incomplete taxonomic resolution (Piñol et al., 2015). For example, Tang et al. (2015) demonstrated that mapping shotgun-sequenced data to complete mitogenomes improved identification and quantitation of species in bee mock communities. Mitochondrial metagenomics can also harness more complete mitogenome sequences to infer phylogenetic relationships from bulk samples (Andújar et al., 2015; Crampton-Platt et al., 2015).

While advancements in sequencing technology have made it feasible to generate the voluminous data on which environmental mitogenomics depend, the lack of well-curated mitogenomic sequence databases presents a bottleneck (Arulandhu et al., 2017). Environmental mitogenomics depend on these databases because they allow matching of any mtDNA fragment to complete, taxonomically verified mitogenomes (Fig 1b) (Crampton-Platt et al., 2016; Deiner, Renshaw, et al., 2017). When such databases exist, any recovered fragment can

improve inference about abundance (Braukmann et al., 2019) and yield valuable information on species presence, which together with high spatial resolution can give you distributional information.

Existing genetic information in public reference databases can facilitate assay design, but issues with data collection make them potentially unreliable. GenBank® (Benson, 1996; Clark et al., 2016) cannot fill the need for curated reference databases because GenBank's sequence data is not uniformly linked to taxonomically verified vouchers. The lack of vouchers weakens the link between DNA sequence and taxonomic identity and prevents independent verification (Meiklejohn et al., 2019). In addition, GenBank does not always screen for contamination or check for errors in protein coding sequences. Quality-checking at GenBank has improved (Leray et al., 2019), but the sequence data it holds can still be of draft quality, contain errors, and be incorrectly assigned taxonomically (Meiklejohn et al., 2019), particularly at the species or subspecies level (Locatelli et al., 2020). Identification errors may particularly plague speciose invertebrate groups (Leray et al., 2020). RefSeq (O'Leary et al., 2016), GenBank's curated and well-annotated sequence dataset, solves some of these problems by incorporating additional rounds of error-checking and provides information for the entire mitogenome. However, it is far from comprehensive—when this study began, RefSeq contained sequence data for just 44% of Oregon's freshwater fish species and much of these data were derived from specimens collected outside the state.

The Barcode of Life Data System (BOLD) provides an alternative to GenBank with more rigorous voucher requirements and consistent use of validation via translation to detect pseudogenes, contaminant checking, and other tools to identify data anomalies and low-quality records (Ratnasingham & Hebert, 2007). However, BOLD skews heavily to information from Cytochrome c oxidase I (COI) due to BOLD's initial development around a single >500 bp barcode region in that gene. As of this writing, COI sequences represent 80.8% of the chordate data available and 82.4% for ray-finned fishes. Although remarkably diagnostic for many species, COI markers often fail to discern recently diverged sister species pairs and may fail to amplify certain taxa due to poorly

conserved primer-binding regions (Deagle et al., 2014). In some cases, other mitochondrial regions contain better conserved sites across taxa for primer placement and do a comparable job of distinguishing species. For example, Miya et al. (2015) found that two 20-30 bp conserved regions flanking a hypervariable region of the 12S mitochondrial gene provided the most suitable fish metabarcoding primers.

Reference databases also require denser intraspecific sampling to identify loci that best diagnose species or geographically structured variation within species. Without multiple sequences for each species, researchers cannot test primer-binding specificity and species diagnosability *in silico*. Here again, available reference sequence databases fall short. GenBank's curated RefSeq database, for example, is non-redundant by design (About RefSeq 2021), with each species associated with only one complete mitogenome. Overall, the data gaps associated with GenBank and BOLD introduce uncertainty and potential error into the eDNA assay design process, making sole reliance on these resources for sequence data problematic.

Regional databases have the benefit of being able to curate full mitogenomic data, providing sequence information for genes and intergenic regions, control error-checking, and identify and resolve taxonomic/genetic inconsistencies through re-sampling and re-validation (Astrin et al., 2013; Buckner et al., 2021; de Santana et al., 2021). The ideal option for developing management-quality eDNA biodiversity surveys would involve extending the "BOLD model" to create curated reference databases of mitogenome sequences tied to vouchered specimens collected throughout discrete regions. Langlois et al. (2021) echoed this need to expand the range of species with full mitogenomic sequence information. The authors specifically call for full mitochondrial genome sequences for multiple examples per species so that robust, comprehensive sequence alignments can support the development of assays that avoid cross-binding of primers to non-target taxa or non-binding of primers to target DNA (Langlois et al., 2021). In addition to improving qPCR primer design, databases of full mitogenomic sequence information provide the data needed to correctly assign metabarcoding sequence queries to taxa regardless of the primer binding site in the mitogenome.

It has become practical and affordable enough for a small consortium to sequence and assemble hundreds of mitogenomes using a single Illumina Novaseq sequencing lane. This means that little impedes development of the curated mitogenomic reference sequence databases needed to prepare for PCR-free mitogenomics, and to develop, test, and query single-species and metabarcoding eDNA assays.

Here, we provide a roadmap for constructing such a curated mitogenomic reference library and evaluate the resulting data to determine the effectiveness of subsections of the mitogenome and the organelle in its entirety to identify species. This effort was motivated by the Oregon Biodiversity Genome Project (OBGP; www.obgp.org), a multi-institution collaboration between scientists and managers at Oregon State University, the Oregon Department of Fish and Wildlife, and the United States Forest Service. The OBGP aims to develop a regional genetic reference database to facilitate statewide eDNA monitoring programs for Oregon's resident freshwater fishes. The specific goals of the OBGP, as outlined in our roadmap (Fig 2), are to: (1) use sterile laboratory methods to collect 10 georeferenced full-bodied vouchers of each freshwater fish species from dispersed watersheds in Oregon; (2) archive and link voucher specimens, tissues, and metadata for taxonomic verification and revision; (3) sequence full mitogenomes from multiple specimens per species; and (4) make all curated data publicly available. While biodiversity and geographic complexity differs from region to region, this study provides a realistic sense of the effort needed to construct a database covering ~150 species spread across ~250,000 km^2. By curating this reference database of full mitogenomes, we created the taxonomic reference information needed to identify freshwater, anadromous, and estuarine fish species found in Oregon and bordering states by any mitogenome-based single-species eDNA or metabarcoding assay and set the stage for future PCR-free environmental mitogenomics methods. Our approach also analyzes the efficacy of various regions of the mitogenome to identify species and provides pipelines that can guide other organizations as they develop reference sequence databases for their taxa and regions of interest.

Materials and Methods

Voucher Specimen and Tissue Collection

The study area initially encompassed the state of Oregon—the region of interest for our eDNA monitoring program—and expanded to a few sites in northern California and Washington State (Fig 3). To strategize sample collection, we examined historical location records in fish collections such as the Oregon State Ichthyology Collection and conferred with local biologists to identify current distributions. For cases where we knew or suspected that deeply divergent evolutionary lineages existed in the present concept of a species, we aimed to include representatives of all lineages. We ultimately identified 146 native and nonnative freshwater fish species and lineages that are currently found in Oregon and strategized collections to span watersheds throughout the state (Appendix S1).

To facilitate consistent sampling, we provided sampling kits (Appendix S2 Box S1) to collectors that contained a 500-mL Nalgene bottle filled with 10% formalin, a 2.0 mL cryotube filled with 95% EtOH, a sterile scalpel, scissors and tweezers, a bleach wipe, latex gloves, a detailed sampling protocol to ensure consistent tissue sampling and data collection (Appendix S2 Box S2), and a field notes sheet (Appendix S2 Box S3) for metadata collection. Collectors anesthetized and euthanized all fish specimens prior to tissue collection by immersion in an aqueous solution of Tricaine mesylate (MS-222) (400 mg MS-222, 400 mg sodium bicarbonate, 1 L water). We instructed all partners to collect a minimum of ~0.5 cm^3 of tissue from each specimen, which was then placed in 95% EtOH for DNA extraction and sequencing. Euthanized fish were placed in 10% formalin as voucher specimens, thereby ensuring preservation of diagnostic features. The cost outlay for enough kit components to collect 1500 individuals totaled US$16,185.20 (Appendix S6).

Taxonomic Verification, Accession, and Cataloging

Fish biologists identified specimens provisionally in the field and then Oregon State Ichthyology Collection taxonomists verified or refined those identifications by morphological examination and reference to published

keys (Markle & Tomelleri, 2016; Wydoski & Whitney, 2003). The Oregon State Ichthyology Collection is in the process of accessioning and cataloging all vouchers and tissues. During that process, the curators input the metadata associated with each specimen and collection event into a relational database, and the full-bodied voucher specimens are transferred from formalin to isopropyl alcohol for permanent storage in a dedicated collection facility that complies with modern fire and earthquake safety codes. Tissues are stored in 2.0 mL cryotubes in -80°C freezers. Total expenditures on storage supplies such as jars, lids and preservation fluid came to US$8,624.21 (Appendix S6).

Mitogenome Assembly

To capture geographic genetic variation of each resident species across its distribution in Oregon, we sequenced the first collected representative of each species and subsequently sequenced specimens collected from separate watersheds, when possible. We stored gzipped fastq sequencing files on 2 x 1TB enterprise NL-SAS hard drives and performed mitogenome assemblies on 4 x 2.30 GHz 16-core processors using 512GB ECC RAM (Total hardware cost $12,368.33; Appendix S6). Mitochondrial genomes were assembled *de novo* from raw paired reads using SPAdes assembler initially (versions 3.12.0-3.15.3) (Bankevich et al., 2012) and getOrganelle 1.6.2 or 1.7.5 (Jin et al., 2020) once released. We annotated all mitochondrial sequences using a combination of MITOS^2 WebServer (Al Arab et al., 2017; Donath et al., 2019) and Geneious 10.2.6 using annotations from identical or closely-related species.

Mitogenome Variability

To analyze intra- and interspecies mitogenome variability, assembled mitogenomes from each species were aligned with MUSCLE (Edgar, 2004) in Geneious 10.2.6 using default parameters. After reciprocal rounds of morphological examination and molecular clustering, we aligned sequences of species from within the same family and then aligned these family clusters to create a master alignment of all sequences. To identify taxonomically diagnostic regions for efficient eDNA assay development, we used the R package SPIDER (Brown

et al., 2012) to perform a sliding window analysis on the master alignment to locate areas with the highest density of taxonomically diagnostic nucleotides (TDN)—defined as locations where a nucleotide is fixed within species and different or unaligned in all other species. To identify genes with high variability, we plotted variability with heat maps, parallel coordinate plots, and radar charts using the R packages Superheat (Barter & Yu, 2018), GGally (Schloerke et al., 2021), and fmsb (Nakazawa, 2021) respectively. Gene regions <690 base pairs in length—ATP6, ATP8, NAD3, NAD4l, NAD6, and all tRNA genes—were not included in our analyses of individual genes. We treated described subspecies as full species for the purpose of calculating mean percent identities. Calculation of intraspecies, intrafamily/interspecies, and interfamily/interspecies mitogenome identities and the proportional relationships among these identities required mitogenomes of multiple specimens for each species within a family. Seven families satisfied these requirements: Catostomidae, Centrarchidae, Cottidae, Cyprinidae, Ictaluridae, Petromyzontidae, and Salmonidae.

Utility of Different Mitogenomic Regions for Species Identification

We evaluated the relative success of subregions of the mitogenome to identify fish species by first extracting subsets from the alignment of 313 mitogenomes. Using Geneious 10.2.6, we created miFish (Miya M. et al., 2015) and Teleo (Valentini et al., 2016) amplicons with Primer3 (Untergasser et al., 2012), and extracted 12S, 16S, NAD2, CO1, NAD4, NAD5, and D-loop regions from the alignment using annotations as guidelines. We examined the entire mitogenome along with regions spanning the 12S, trnV, and 16S genes and the NAD4, trnH, trnS1, trnL1, and NAD5 genes. We performed BLASTn queries of these regions against a local database created from our 313 mitogenomes. We parsed the results from this BLASTn query to determine the effectiveness of the different regions to successfully identify species.

Data Sharing

Mitogenome data generated for this project have been deposited in GenBank under the Oregon Biodiversity Genome Project BioProject. GenBank accession numbers and sequence data are included in Supplemental

Information (Appendix S3 columns: genbank_accession, sequence respectively; Mitogenome FASTAs are available in Appendix S7). Sequence data are also available at www.obgp.org/downloads. As of the time of this writing, linked voucher, tissue, and DNA extract accessioning into the Oregon State Ichthyology Collection was ongoing. The voucher specimens accessioned and cataloged to date are available and searchable via webportal at https://webportal.specifycloud.org/osichthyology/, which the collection's curators update periodically.

Further details of Voucher specimen and tissue collection, Taxonomic verification, accession, and cataloging, Mitogenome assembly, and Utility of different mitogenomic regions for species identification are available in the Supplemental Material (Appendix S8). Information regarding DNA extraction and sequencing is solely available in the Supplemental Material (Appendix S8).

Results

Voucher Specimen and Tissue Collection

We collected 625 specimens representing 128 fish species or distinct lineages within species complexes. These specimens originated from more than 240 localities in Oregon. Twelve additional tissue samples of four species were acquired from natural history collections, bringing the total number of species represented in the database to 132. Of these 132 species, 119 represent the original 146 fish species identified by Oregon Department of Fish and Wildlife as native or naturalized in Oregon. The remaining 13 species belong to 11 coastal estuarine species not included in our initial freshwater collection plan, plus one species endemic to western Washington (Olympic Mudminnow, *Novumbra hubbsi*) and one newly identified lineage of Paiute Sculpin (*Cottus beldingii* ssp.) from the John Day River Basin in central Oregon.

Taxonomic Verification

After clustering sequences based on maximum-likelihood (ML) inference and examining voucher specimens in the lab, we refined or corrected 31 field identifications (9.9%) (Appendix S3, taxonomic_assessments; Appendix

S2, Figures S1-S5). Specimens from Cottidae represented the highest proportion of conflicting identifications (n=16; 52%) likely because the group is the subject of active taxonomic research to identify the geographic, genetic, and anatomical limits of their component evolutionary lineages (M. K. Young et al., 2022). The most current morphological identification keys to *Cottus* do not reliably separate all the species and species complexes in this genus.

Of the 31 total refinements, fifteen (48%) (Cottidae, n=12; Catostomidae, n=1; Petromyzontidae, n=1; Ictaluridae, n=1) had ambiguous or conflicting taxonomic assessments and were assigned to a species or species complex based on ML inference. Of the cottids, nine resided in the *Cottus gulosus/perplexus complex* and were assigned to either *C. perplexus* (n=8) or *C. gulosus* (n=1), while OBGP-2017-269 resided in the *Cottus asper/perplexus* and was assigned to *C. asper*. OBGP-2017-148 matched some but not all the features normally diagnosing *C. beldingii* and clustered with the newly identified *C. beldingii* lineage, and OBGP-2018-012 was assigned to *C. confusus* as this was suggested by ML inference and supported by the morphological identification key for the species. Of the three remaining specimens with ambiguous taxonomic assignments, the sucker specimen OBGP-2018-023 clustered with *Catostomus bondi* yet lacked all the diagnostic features of this species and was assumed to be a hybrid with *Catostomus columbianus*. OBGP-2016-007 was a transforming lamprey microphthalmia with caudal fin pigmentation suggesting *Entosphenus lethophagus*, and a catfish specimen OBGP-2018-046 that had some of the diagnostic features of both *Ameiurus natalis* and *A. nebulosus* but clustered with *A. nebulosus* and was assumed to be a hybrid.

Morphological assignment superseded ML clustering for four individuals (13%) consisting of three cottids (OBGP-2017-273, OBGP-2017-346, OBGP-2019-178) and one catostomid (OBGP-2017-101). ML clustering disagreed with in-field assignment and was concordant with morphological assessment for six specimens (19%). These consisted of challenging-to-identify species from Centrarchidae (OBGP-2017-308, OBGP-2017-381, OBGP-2018-159) and Petromyzontidae (OBGP-2017-248, OBGP-2017-250), along with a single cyprinid that

was a xanthic morph of *Pimephales promelas* (OBGP-2018-033). Three vouchers (10%), all centrarchids, were of poor quality and made morphological identification challenging (OBGP-2017-275, OBGP-2017-360, OBGP-2019-057). An additional two catostomids (7%) (OBGP-2018-006, OBGP-2018-092) appears to have been switched during processing, and one cottid (3%) was erroneously noted to be a subspecies of *Cottus klamathensis* (OBGP-2017-218) during wet lab processing.

Mitogenome Sequencing and Assembly

In total, 313 assembled mitogenomes representing 128 collected species and lineages were used for downstream analysis (Table 1). Nearly all *de novo* assemblies (96.8%; n = 303) resolved as a single mitochondrial contig with an overlapping splice point. The remaining assemblies were derived from either: (a) multiple contigs with overlapping splice points (n = 3); (b) a single contig with a nonoverlapping splice point in an intergenic area with mononucleotide C repeats (n = 6); or (c) multiple contigs from different SPAdes runs with overlapping splice points (n = 1). All mitogenomes had GC content between 38.90% and 49.50% (mean 45.14%) except for within Petromyzontidae, where GC content ranged from 37.90% to 38.70% (mean 38.05%) (Appendix S3 gc_content). Mitogenome sizes ranged from 16098 to 17185 bp in length (mean 16590). All but 17 assembled mitogenomes had error-free contigs when measured with k=31 using Merqury. Assembled mitogenomes with errors had QVs between 40.7507 and 57.0952 (Appendix S3 contig_qv) indicating errors in the range of 1 in ~10,000 bp to 1 in ~1,000,000 bp, respectively. Read mapping showed anomalous coverage in intergenic regions of 36 assemblies. These anomalies were generally located in areas with nucleotide repeats and manifested as either dips or spikes in coverage. Dips were likely due to sequencing and/or assembly errors and spikes were assumed to be the result of nuclear mitochondrial DNA. These anomalies were not sufficient to exclude mitogenomes from downstream analyses (Appendix S3 assembly_notes).

Mitogenome Variability

The sliding window analysis of our alignment of 313 complete mitogenomes revealed that mean taxonomically diagnostic nucleotides per 150-base window shifted at 20-base intervals (TDN/$w_{150}i_{20}$) in analyzed gene regions were as follows: COI 7.257, CytB 9.451, NAD1 13.381, NAD5 17.092, NAD4 18.226, NAD2 20.065, 12S 20.814, 16S 25.726 (Fig 4b). The highest concentrations of TDNs occurred in the D-loop and the intergenic region between the NAD2 and COI genes (Fig 4a). Sliding window analyses of aligned taxonomic subsets of the entire mitogenome for catostomids, centrarchids, cottids, cyprinids, ictalurids, petromyzontids, and salmonids suggested that the density of TDN varies by taxonomic group (Fig 5) with mean TDN/$w_{150}i_{20}$ of 10.656, 12.104, 18.871, 19.316, 20.209, 21.024, and 26.082 for these families, respectively.

We used heat maps (Fig 6a) to illustrate the degrees of similarity between families in different gene regions, although all genes have diverged sufficiently to diagnose familial lineages. These results showed that sequence identity among families in the COI gene exceeded that of other coding and noncoding gene regions (Fig 6a) suggesting that COI is the mitogenomic region in which families differ the least, a conclusion that concords with the results from the sliding window analysis. NAD2, NAD5, and 16S contrasted the most among families and species in overall percent identity. Despite different degrees of intrafamily identity in the mitochondrial regions we examined, our analysis suggested that divergence in all mitochondrial genes and the D-loop is sufficient to identify taxa at the family level (Fig 6a).

Zooming in to species differences within families, after errors in field identification were corrected, all species resembled members of their own species in mtDNA sequences more than they resembled members of other species within the same family, as expected (Appendix S4 Table S1). Species within the same family differed most in the percent identity of the D-loop (0.855), followed by the NAD2 (mean 0.885), NAD5 (0.896), and NAD4 (0.896) genes. The 12S (0.965) and 16S (0.957) genes were the least differentiable regions (Fig 7). Intraspecies mean percent identities for the 12S and 16S rRNA genes and all coding genes >690 bp ranged from 98.259 to 99.975% and from 92.428 to 98.249% for the D-loop (Appendix S4 Table S2). This comparison

illustrates that the D-loop is less conserved within species than are any of the other gene regions (Fig 8). The most conserved genes were 12S, 16S, and COX2, with the lowest mean values found in the NAD2 gene nevertheless still exceeding 99% identity (Appendix S4 Table S2). Radar charts of mean percent intraspecies, intrafamily interspecies, and interfamily interspecies identity (Fig 8) illustrated that different genes are more conserved among species within certain families than others. For example, species in Catostomidae varied little in sequences from rRNA and all three COX genes, while the 12S and 16S genes were fairly conserved among salmonid and cottid species. Non-rRNA regions in Salmonidae and Cottidae, and all gene regions and the D-loop in Cyprinidae, Centrarchidae, and Ictaluridae contained diverged interspecies sequences. Full mitogenomes were highly conserved within species (mean 99.493% identity) and had sufficient divergence among species in the same family to suggest they would be diagnostic at the species level for Oregon fishes (Fig 9).

Utility of Mitogenomic Regions for Species Identification

Plots of parsed results from BLASTn queries suggested that the full mitogenome, concatenated gene regions from the mitogenome, and the NAD5 gene are superior to miFish and Teleo metabarcoding primers and most individual mitochondrial genes for producing first hits that match specimens to species and to described subspecies (Fig 10). However, queries using every gene region and Teleo and miFish primer amplicons do a reasonable job of producing first hits matching some specimens to the species or subspecies level. Queries of whole gene regions performed better overall than queries of shorter amplicons.

Discussion

A Blueprint for Constructing Mitogenomic Databases

We demonstrated a robust, affordable, and feasible blueprint for constructing mitogenomic databases. The workflow begins with the collection of reference specimens and progresses through taxonomic verification, permanent accessioning of specimens, tissues, and DNA, mitogenome assembly, and open-source provisioning

of complete mitogenomes. Such databases can help to refine the taxonomy of understudied or difficult groups, guide the discovery and delineation of cryptic species or distinct population segments, and facilitate the transition to eDNA-based monitoring of aquatic biodiversity.

Though beginning such a project can seem daunting, the steps for carrying out a similar endeavor are straightforward: 1. Using historical collection data and local knowledge, determine all focal species and their distributions, 2. Break up the region of interest into manageable subregions for sampling, 3. Create a sampling plan to collect 3-10 individuals per species/lineages of interest and begin the sampling effort using accepted standards for metadata collection (Rimet et al., 2021), acquiring tissues from vouchered specimens in natural history collections whenever possible, 4. Sequence and assemble specimens as they accumulate, measuring intraspecies sequence variability to inform continued collection.

We have cataloged the collection strategy and wet and dry laboratory pipelines we used for our bottom-up development of an eDNA biodiversity reference collection and sequence database and provide an easy-to-follow roadmap (Fig 2). This bottom-up approach harnesses the expertise, knowledge, and resources of researchers and managers within their region and taxa of interest, an essential strategy as these individuals possess the intimate knowledge of species, taxonomy, and geography needed to plan expeditions, carry out collections, and identify specimens.

Because projects of this scale require moderate financial support and substantial human effort, we strongly recommend assessing available resources before launching a new endeavor. We were able to complete this project on a relatively low budget (Estimated cost without labor ~$250 per mitogenome; Appendix S6) because individuals donated considerable amounts of their time collecting throughout the state of Oregon and because collaborating institutions provided us with genetic laboratory facilities and sequencing at reduced costs. The workflow also depends on taxonomic expertise in identifying specimens within difficult families, namely those

featuring many morphologically similar species or undescribed cryptic species. Finally, access to the infrastructure and archival capacity of a natural history collection is vital because the voucher specimens must be cataloged properly and preserved in perpetuity for the science to be verifiable and repeatable (Astrin et al., 2013; Buckner et al., 2021; Prendini et al., 2002).

Researchers seeking to construct reference libraries for non-piscine taxa should first understand the structure and makeup of the mitochondrial and nuclear genomes of those groups prior to curating mitogenomic sequences. In particular, they should ensure that the proposed wet and dry laboratory pipelines can successfully resolve mitogenomes in those groups. Fish mitogenomes contain fewer repeats, insertions, and deletions than those of other vertebrates (Formenti et al., 2021), and all these can cause problems in the sequencing and assembly pipeline (Tørresen et al., 2019). Increased sequencing depth may be required in other taxa if using short read sequencers alone. Alternatively, a combination of long-read and short-read sequencing may readily resolve complete mitogenomes (Formenti et al., 2021).

Taxonomically Informative Genes for Fishes

Our reference sequence database provides a valuable genetic resource for analyzing mitochondrial genetic variability among Oregon's freshwater fishes and gauging capacity to identify species in eDNA assays. Our analyses illustrate that mitochondrial sequences at every level, from individual genes to the entire mitogenome, are sufficiently conserved within species to provide reliable identifications. However, not all mtDNA regions are equally good at distinguishing between closely related species and certain complete genes and concatenated gene regions (such as the NAD regions) are generally superior (Fig 10). Sequences must have diverged sufficiently among taxa to discern them, so longer regions and faster evolving sections of the mitogenome are preferable for diagnosing recently separated lineages.

Sections of the COI gene have been recommended for barcoding animals (Hebert, Cywinska, et al., 2003) and metabarcoding eukaryotes (Meusnier et al., 2008) and metazoans (Andújar et al., 2018), but they may not be optimal for identifying certain taxa, especially when only short stretches of DNA are retrievable. For certain taxonomic groups COI sequences do not consistently cluster or associate sequences with their assigned species when barcoding (Waugh, 2007) or metabarcoding (Collins et al., 2019). Previous analyses by Hebert, Cywinska et al. (2003) and Hebert, Ratnasingham et al. (2003) examined the use of COI for species identification in Lepidoptera by quantifying sequence divergence and using NJ analyses and multidimensional scaling to assign sequences to species. They extrapolated the general suitability of their COI barcode to diagnose animal species from their success in these taxonomically narrow trials. Though their arguments in favor of the COI as the core of a global bioidentification system for animals were logical, they were also speculative (Hebert, Ratnasingham, et al., 2003). They did not assess the comparative merits of the COI over other mitochondrial genes and explicitly stated the need to validate the diagnosability of the COI gene for different taxonomic groups (Hebert, Cywinska, et al., 2003). This has been done for the COI barcode for a variety of taxonomic groups over the intervening decades (invertebrates: (Cywinska et al., 2006; Sheffield et al., 2009; M. R. Young et al., 2019); fish: (Zemlak et al., 2009); birds: (Hebert et al., 2004; Kerr et al., 2009); amphibians: (Smith et al., 2008); mammals: (Francis et al., 2010)) with results based on sequence divergence and NJ clustering analyses suggesting that, for arthropods and vertebrates, this barcode is useful for parsing these groups taxonomically. It is unclear, however, to what degree the COI is taxonomically diagnostic to the species level for all metazoans. Relatively low percent identity between multiple species does not necessarily equate to species-level diagnosticity (Appendix S4, Fig S9). A more complete evaluation of the comparative diagnosability of different parts of the mitogenome is therefore needed for a broad range of taxa. Here, we demonstrate that for Oregon's freshwater fishes, regions other than the COI, in particular the NAD regions, have the potential to better identify species in certain taxa.

A variety of factors contribute to how successfully a region parses the independently evolving lineages we typically call species. Areas with relatively high interspecies genetic distance and concentrations of

taxonomically diagnostic nucleotides within families are likely candidates for diagnosing to the species level, whether barcoding, metabarcoding, or performing single-species detection. To "capture" these regions for single-species qPCR assays the goal is to capture eDNA from the target species and no other species, so areas with high intrafamily distance, high intraspecies identity, and high mean concentrations of TDNs are likely the best candidates. Unlike single-species qPCR assays, metabarcoding primers need to capture eDNA from a broad range of taxa—different families, orders, classes, or even phyla—so there need to be shared regions (typically between 18 and 27 bases long) that can permit primer binding and avoid species dropout. Essentially, a "Goldilocks" zone is needed for metabarcoding: a region with sufficient genetic divergence to distinguish between species, but not to the degree that shared regions are unavailable for primer binding. For this reason, the hairpin-loop structure of both rRNA regions makes them appropriate for metabarcoding and explains why the most referenced fish metabarcoding primers are found in the 12S region (Miya M. et al., 2015) despite this region's high within-family interspecies percent identity relative to other regions.

For Oregon's freshwater fishes, we found multiple gene regions and the D-loop had high interspecies genetic distance and concentrations of TDNs within families, suggesting there are numerous alternatives to the COI gene for species identification using eDNA. For single species analyses, the results were mixed depending on the family examined. Salmonidae, Cyprinidae, Cottidae, and Petromyzontidae had highest mean TDN/$w_{150}i_{20}$ in the D-loop (Table 2; Fig 5), although this region also has low relative intraspecies identity compared to other mitochondrial regions (Fig 8). Regions with a combination of all three important factors—high intraspecies identity, high intrafamily distance, and high mean TDN/$w_{150}i_{20}$—were the NAD2 gene for Salmonidae, Centrarchidae, Cottidae and Petromyzontidae, the NAD5 gene for Cyprinidae, and the NAD4 gene for Catostomidae and Ictaluridae (Table 2; Fig 7; Table S2). The 12S and 16S regions contain the TDN ‘spikes’ (Table 2) we would expect to see given that loop regions permit the introduction of mutations while the complementary hairpin regions around them are conserved. These clusters of TDNs sandwiched by conserved hairpin areas illustrate why these rRNA regions are better candidates for metabarcoding primers than the COI.

Although the 16S rRNA region had these "spikes" across more families than other genes (Table 2), our analysis of diagnostic nucleotides suggests that for metabarcoding, no single gene region is clearly best for all families of resident freshwater fish in Oregon, and primer-binding requirements further restrict which areas can be used.

To ultimately be of use for metabarcoding, a region needs to diagnose species effectively and consistently and be sequenceable using contemporary technologies. Our comparative BLAST analysis provides insight into how consistently various regions of the mitogenome diagnose species or subspecies (Fig 10). The results suggest that, when performing BLASTn sequence queries—the standard query used for eDNA metabarcoding results—multiple genes and the full mitogenome are generally more successful at identification than most individual genes and mini-barcodes. The exception is the NAD5 gene, which produced the most target-specific results along with queries using the full mitogenome and concatenated NAD4/NAD5 or 12S/16S regions. This suggests that the NAD5 gene and certain concatenated genes within the mitogenome may be useful for single species assays and with future PCR-free based approaches. However, the low interspecies percent identity of the NAD5 and other NAD genes (Fig 8) may prevent locating non-degenerate multispecies metabarcoding primers that capture regions short enough to be sequenced using Illumina technologies. In addition, complete single and concatenated genes are too long to be sequenced using these next generation methods. For example, the ~650 bp COI region published by Hebert, Ratnasingham et al. (2003) exceeds the length feasible for Illumina high throughput sequencing (Meusnier et al., 2008), so subsections need to be used for metabarcoding. However, capturing regions of the COI that are sufficiently short for Illumina sequencing often requires the use of primers that are highly degenerate (Collins et al., 2019; Deagle et al., 2014). Due to having a wide range of optimal primer melting temperatures, degenerate primers may differentially amplify target species, of particular concern when targeting extraorganismal eDNA (Hajibabaei et al., 2019). The 16S gene may present a potential alternative to 12S mini-barcode regions for eDNA metabarcoding of fishes as it identifies targets to species or subspecies relatively well (Fig 10) and has high interspecies percent identity (Fig 8) and TDN spikes (Table 2). Additional studies evaluating subsections of the 16S gene for this purpose would be informative.

In cases where no single barcode can separate all targeted species, several viable options exist. For single-species and metabarcoding assays that rely on short barcode regions, it may be necessary to use multiple regions for metabarcoding or perhaps a diagnostic region in the nuclear genome such as the ITS1 gene to discern closely related congeners (Dysthe et al., 2018). Having more extensive mitogenomic data is necessary for multilocus metabarcoding, and for PCR-free approaches the significant intrafamily interspecies variability of full mitogenomes permits improved species identification over single genes, even among relatively conserved taxa. The benefits of using full mitogenomes or a combination of strategically valuable mitochondrial genes derived from full mitogenomes to discern Oregon's resident fish species is illustrated by our analysis of *in silico* BLASTn queries (Fig 10).

Biological Challenges in Distinguishing Species

It is important to note that the failure to correctly identify the presence of a species using eDNA may be the biological reality rather than a fault of the method. Species that often confound taxonomists and are difficult to distinguish morphologically also appear to be difficult to resolve genetically; not even even full, reliable, mitogenomic sequences may be able to resolve such cases. For example, hybridization and organelle introgression from secondary contact can obscure the relationships of different species in an environment (Dowling et al., 2016; Forsythe et al., 2020). Due to being inherited matrilineally, mitogenomic information on its own cannot distinguish hybrid species, and nuclear genetic information may be needed to untangle the genetic complexities of introgression resulting from secondary contact and hybridization—the likely culprit behind difficulties identifying certain catostomid species (Dowling et al., 2016). Difficulties with cottid identification may result from insufficiently diverged lineages or widely variable morphology (Rowsey & Egge, 2017), and the tendency of lamprey taxa to rapidly derive non-parasitic "satellite species" from parasitic species in sympatry (Salewski, 2003; Vladykov & Kott, 1979) often leaves insufficient time for genetic divergence to accrue (April et al., 2011; Brownstein & Near, 2023). Understanding these shortcomings is essential when

undertaking biodiversity assessments using eDNA, and even though the method may not be perfect, the ability to assess a broad variety of taxa using noninvasively collected environmental samples is profoundly transformative.

Mitogenomes in a Nuclear Future

Despite the challenges involved with a project of this scale and the limitations of mitogenomes to genetically resolve hybrids and certain closely related species, full mitogenomic data provide both a useful genetic reference for species identification and the genetic information needed to develop primers for single-species and metabarcoding assays. It also furnishes researchers with the data needed to move away from microgenomics such as barcoding or metabarcoding and into capture enrichment (Wilcox et al., 2018) or PCR-free environmental genomics, which by definition solve the problems associated with PCR amplification biases (Piñol et al., 2015). Such methods make accurate quantification of relative species abundance in a sample a real possibility (Yang et al., 2021). Full mitogenomic data also frequently diagnose individuals to species more reliably than do shorter sequences (Fig 10). Compiling regional mitogenome databases expands the global repository of available genetic data and builds local capacity to query metabarcoding reads and develop or select the qPCR and metabarcoding primers that best support eDNA-based monitoring of biodiversity.

In anticipation of future applications, many organizations are assembling whole nuclear genomes. Example consortia include the “Bat1K” (Teeling et al., 2018) and “1000 Fungal Genomes” (Grigoriev et al., 2014) projects, sequencing 1000 species of bats and fungi respectively, the “i5k” consortium sequencing 5000 arthropod genomes (i5K Consortium, 2013), the “10KP” and “P10K” projects sequencing 10,000 plant (Cheng et al., 2018) and protist (Miao et al., 2020) species, and the “GIGA” project (GIGA Community of Scientists, 2014) dedicated to sequencing invertebrate genomes. Even more ambitious projects are efforts to sequence genomes from representatives of every vertebrate species (Vertebrate Genomes Project (VGP)) (Rhie et al., 2021), and all of Earth's eukaryotic biodiversity by 2027 (Earth Biogenome Project) (Lewin et al., 2018). These large-scale,

expensive, and top-down efforts will create nuclear genome reference databases for many species in the coming decades, but their global focus is unlikely to provide comprehensive genetic information for any specific geographic region or taxonomic group of interest. For example, as of the time of this writing the VGP has produced 110 nuclear assemblies and published data for 125 mitogenomes (Formenti et al., 2021). This is an invaluable contribution but lacks redundant sequence information. An exception to these large-scale projects is the California Conservation Genomics Project (Shaffer et al., 2022), a regional effort to create nuclear reference genomes and multiple resequenced nuclear genomes for numerous species within the state of California given a budget of US$12M. This presents an exemplar for researchers to pursue similar efforts, and if the affordability and accuracy promised by PacBio for its new generation of long read sequencers becomes a reality, projects of this scale will effectively become democratized. While we should strive to create reference databases of full nuclear genomes for all organisms on the planet, nuclear DNA may not turn up reliably in extraorganismal eDNA from aggregating, non-spawning metazoans (Jensen et al., 2021; Olson et al., 2012) making nuclear reference sequence databases less useful for eDNA applications. There will therefore always be a need for mitogenomic reference data, and much can still be gleaned from complete mitogenomes on their own.

Conclusion

We hope these protocols and insights into mitogenomic variability will encourage researchers around the globe to follow suit and develop their own regional databases and archives of voucher specimens. Widely ranging mitogenomic databases would expand eDNA monitoring potential to more regions. Repositories of vouchered samples and full mitogenomic information as described herein not only provide the genetic information needed to use eDNA effectively for biodiversity studies (de Santana et al., 2021), but also can support investigations of taxonomy, population structure, landscape genetics and multilocus metabarcoding. Coupled with high-molecular-weight DNA extraction for nuclear genome sequencing, a project of this scope grows the global repository of sequence information in anticipation of an environmental genomics future, and lays the

groundwork to compile all the available genetic information for freshwater fishes or other taxa in a region of interest.

Acknowledgements

We thank the Oregon Department of Fish and Wildlife, United States Forest Service Pacific Northwest Research Station, and Oregon State University for financial and logistical support. We are also grateful to the Klamath Tribes, United States Fish and Wildlife Service, and the California Department of Fish and Wildlife for providing valuable contributions toward our collection goals. In addition, we greatly appreciate specimen donations from Mark Buettner, Nolan Banish, Bruce Hansen, Paul Divine, and Dave Hering. Laura Hauck and Lucas Longway provided invaluable assistance with laboratory training and Shawn Clements helped to get this project off the ground. Many thanks to the University of Washington, the University of Alaska Museum of the North, and the Oregon State University Ichthyology Collection for providing us with vouchered tissue samples. We also thank Doug Markle and Álvaro Cortés for providing taxonomic identification assistance. In addition, Tom Stahl, Marc Johnson, and members of the Levi Lab reviewed the manuscript and provided valuable feedback for which we are also grateful.

Author Contributions

ED, RC, BEP, TL, and PK conceived the project. JA, TC, TF, BS, BEP, and SS strategized and carried out specimen collection. ED analyzed the data. ED, TL, BS, RC, BEP, JA, PK, and TF wrote the paper.

References

About RefSeq [Internet]. 2021. Bethesda (MD): National Center for Biotechnology Information; [cited 3 February 2021]. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/

Al Arab, M., Höner zu Siederdissen, C., Tout, K., Sahyoun, A. H., Stadler, P. F., & Bernt, M. (2017). Accurate annotation of protein-coding genes in mitochondrial genomes. *Molecular Phylogenetics and Evolution*, *106*, 209–216. https://doi.org/10.1016/j.ympev.2016.09.024

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjær, K. H., Orlando, L., Gilbert, M. T. P., & Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, *21*(8), 1966–1979. https://doi.org/10.1111/j.1365-294X.2011.05261.x

Andújar, C., Arribas, P., Ruzicka, F., Crampton-Platt, A., Timmermans, M. J. T. N., & Vogler, A. P. (2015). Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology*, *24*(14), 3603–3617. https://doi.org/10.1111/mec.13195

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*(20), 3968–3975. https://doi.org/10.1111/mec.14844

April, J., Mayden, R. L., Hanner, R. H., & Bernatchez, L. (2011). Genetic calibration of species diversity among North America's freshwater fishes. *Proceedings of the National Academy of Sciences*, *108*(26), 10602–10607. https://doi.org/10.1073/pnas.1016437108

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F. B., Barreto Crespo, M. T., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hochegger, R., Klingl, A., Lundberg, L., Natale, C., … Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, *6*(10). https://doi.org/10.1093/gigascience/gix080

Astrin, J., Zhou, X., & Misof, B. (2013). The importance of biobanking in molecular taxonomy, with proposed definitions for vouchers in a molecular context. *ZooKeys*, *365*, 67–70. https://doi.org/10.3897/zookeys.365.5875

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barter, R. L., & Yu, B. (2018). Superheat: An R Package for creating beautiful and extendable heatmaps for visualizing complex data. *Journal of Computational and Graphical Statistics*, *27*(4), 910–922. https://doi.org/10.1080/10618600.2018.1473780

Béjà, O. (2004). To BAC or not to BAC: Marine ecogenomics. *Current Opinion in Biotechnology*, *15*(3), 187–190. https://doi.org/10.1016/j.copbio.2004.03.005

Benson, D. (1996). GenBank. *Nucleic Acids Research*, *24*(1), 1–5. https://doi.org/10.1093/nar/24.1.1

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bragg, L., & Tyson, G. W. (2014). Metagenomics using next-generation sequencing. In I. T. Paulsen & A. J. Holmes (Eds.), *Environmental Microbiology* (Vol. 1096, pp. 183–201). Humana Press. https://doi.org/10.1007/978-1-62703-712-9_15

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, *19*(3), 711–727. https://doi.org/10.1111/1755-0998.13008

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, *12*(3), 562–565. https://doi.org/10.1111/j.1755-0998.2011.03108.x

Brownstein, C. D., & Near, T. J. (2023). Phylogenetics and the Cenozoic radiation of lampreys. *Current Biology*, *33*(2), 397-404.e3. https://doi.org/10.1016/j.cub.2022.12.018

Bru, D., Martin-Laurent, F., & Philippot, L. (2008). Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Applied and Environmental Microbiology*, *74*(5), 1660–1663. https://doi.org/10.1128/AEM.02403-07

Buckner, J. C., Sanders, R. C., Faircloth, B. C., & Chakrabarty, P. (2021). The critical importance of vouchers in genomics. *ELife*, *10*, e68264. https://doi.org/10.7554/eLife.68264

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P. M., Li, F. W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., & Wong, G. K. S. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*, *7*(3). https://doi.org/10.1093/gigascience/giy013

Clare, E. L., Economou, C. K., Bennett, F. J., Dyer, C. E., Adams, K., McRobie, B., Drinkwater, R., & Littlefair, J. E. (2022). Measuring biodiversity from DNA in the air. *Current Biology*, *32*(3), 693–700. https://doi.org/10.1016/j.cub.2021.11.064

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67–D72. https://doi.org/10.1093/nar/gkv1276

Collins, R. A., Armstrong, K. F., Holyoake, A. J., & Keeling, S. (2013). Something in the water: Biosecurity monitoring of ornamental fish imports using environmental DNA. *Biological Invasions*, *15*(6), 1209–1215. https://doi.org/10.1007/s10530-012-0376-9

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *10*(11), 1985–2001. https://doi.org/10.1111/2041-210X.13276

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*(13), 2937–2958. https://doi.org/10.1111/mec.15472

Crampton-Platt, A., Timmermans, M. J. T. N., Gimmel, M. L., Kutty, S. N., Cockerill, T. D., Vun Khen, C., & Vogler, A. P. (2015). Soup to tree: The phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. *Molecular Biology and Evolution*, *32*(9), 2302–2316. https://doi.org/10.1093/molbev/msv111

Crampton-Platt, A., Yu, D. W., Zhou, X., & Vogler, A. P. (2016). Mitochondrial metagenomics: Letting the genes out of the bottle. *GigaScience*, *5*(1), 15. https://doi.org/10.1186/s13742-016-0120-y

Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 209–230. https://doi.org/10.1146/annurev-ecolsys-110617-062306

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., Barber, P. H., Kraft, N., Wayne, R., & Meyer, R. S. (2019). *Anacapa Toolkit*: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210X.13214

Cywinska, A., Hunter, F. F., & Hebert, P. D. N. (2006). Identifying Canadian mosquito species through DNA barcodes. *Medical and Veterinary Entomology*, *20*(4), 413–424. https://doi.org/10.1111/j.1365-2915.2006.00653.x

de Santana, C. D., Parenti, L. R., Dillman, C. B., Coddington, J. A., Bastos, D. A., Baldwin, C. C., Zuanon, J., Torrente-Vilara, G., Covain, R., Menezes, N. A., Datovo, A., Sado, T., & Miya, M. (2021). The critical role of natural history museums in advancing eDNA for biodiversity studies: A case study with Amazonian fishes. *Scientific Reports*, *11*(1), 18159. https://doi.org/10.1038/s41598-021-97128-3

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*(9), 20140562. https://doi.org/10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J. C., & Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyer belts of biodiversity information. *Nature Communications*, *7*, 12544. https://doi.org/10.1038/ncomms12544

Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., & Pfrender, M. E. (2017). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution*, *8*(12), 1888–1898. https://doi.org/10.1111/2041-210X.12836

Donath, A., Jühling, F., Al-Arab, M., Bernhart, S. H., Reinhardt, F., Stadler, P. F., Middendorf, M., & Bernt, M. (2019). Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research*, *47*(20), 10543–10552. https://doi.org/10.1093/nar/gkz833

Dowling, T. E., Markle, D. F., Tranah, G. J., Carson, E. W., Wagman, D. W., & May, B. P. (2016). Introgressive Hybridization and the evolution of lake-adapted catostomid fishes. *PLOS ONE*, *11*(3), e0149884. https://doi.org/10.1371/journal.pone.0149884

Drummond, A. J., Newcomb, R. D., Buckley, T. R., Xie, D., Dopheide, A., Potter, B. C., Heled, J., Ross, H. A., Tooman, L., Grosser, S., Park, D., Demetras, N. J., Stevens, M. I., Russell, J. C., Anderson, S. H., Carter, A., & Nelson, N. (2015). Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, *4*(1), 46. https://doi.org/10.1186/s13742-015-0086-1

Dysthe, J. C., Franklin, T. W., McKelvey, K. S., Young, M. K., & Schwartz, M. K. (2018). An improved environmental DNA assay for bull trout (*Salvelinus confluentus*) based on the ribosomal internal transcribed spacer I. *PLOS ONE*, *13*(11), e0206851. https://doi.org/10.1371/journal.pone.0206851

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Foran, D. R. (2006). Relative degradation of nuclear and mitochondrial DNA: An experimental approach*. *Journal of Forensic Sciences*, *51*(4), 766–770. https://doi.org/10.1111/j.1556-4029.2006.00176.x

Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Brown, S., Capodiferro, M. R., Al-Ajli, F. O., Ambrosini, R., Houde, P., Koren, S., Oliver, K., Smith, M., Skelton, J., Betteridge, E., Dolucan, J., Corton, C., Bista, I., … Jarvis, E. D. (2021). Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biology*, *22*(1), 120. https://doi.org/10.1186/s13059-021-02336-9

Forsythe, E. S., Nelson, A. D. L., & Beilstein, M. A. (2020). Biased gene retention in the face of introgression obscures species relationships. *Genome Biology and Evolution*, *12*(9), 1646–1663. https://doi.org/10.1093/gbe/evaa149

Francis, C. M., Borisenko, A. V., Ivanova, N. V., Eger, J. L., Lim, B. K., Guillén-Servent, A., Kruskop, S. V., Mackie, I., & Hebert, P. D. N. (2010). The role of DNA barcodes in understanding and conservation of mammal diversity in Southeast Asia. *PLoS ONE*, *5*(9), e12575. https://doi.org/10.1371/journal.pone.0012575

GIGA Community of Scientists. (2014). The Global Invertebrate Genomics Alliance (GIGA): Developing community resources to study diverse invertebrate genomes. *Journal of Heredity*, *105*(1), 1–18. https://doi.org/10.1093/jhered/est084

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., Laramie, M. B., Mahon, A. R., Lance, R. F., Pilliod, D. S., Strickler, K. M., Waits, L. P., Fremier, A. K., Takahara, T., Herder, J. E., & Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307. https://doi.org/10.1111/2041-210X.12595

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., & Shabalov, I. (2014). MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Research*, *42*(D1), D699–D704. https://doi.org/10.1093/nar/gkt1183

Hajibabaei, M., Porter, T. M., Robinson, C. V., Baird, D. J., Shokralla, S., & Wright, M. T. G. (2019). Watered-down biodiversity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples. *PLOS ONE*, *14*(12), e0225409. https://doi.org/10.1371/journal.pone.0225409

Hartmann, N., Reichwald, K., Wittig, I., Dröse, S., Schmeisser, S., Lück, C., Hahn, C., Graf, M., Gausmann, U., Terzibasi, E., Cellerino, A., Ristow, M., Brandt, U., Platzer, M., & Englert, C. (2011). Mitochondrial DNA copy number and function decrease with age in the short-lived fish *Nothobranchius furzeri*: Decline of mitochondrial function in aging fish. *Aging Cell*, *10*(5), 824–831. https://doi.org/10.1111/j.1474-9726.2011.00723.x

Hauck, L. L., Weitemier, K. A., Penaluna, B. E., Garcia, T. S., & Cronn, R. (2019). Casting a broader net: Using microfluidic metagenomics to capture aquatic biodiversity data from diverse taxonomic targets. *Environmental DNA*, *1*(3), 251–267. https://doi.org/10.1002/edn3.26

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: Cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1). https://doi.org/10.1098/rsbl.2003.0025

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, *2*(10), e312. https://doi.org/10.1371/journal.pbio.0020312

i5K Consortium. (2013). The i5K Initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, *104*(5), 595–600. https://doi.org/10.1093/jhered/est050

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., Møller, P. R., & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources*, *21*(3), 690–702. https://doi.org/10.1111/1755-0998.13293

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241. https://doi.org/10.1186/s13059-020-02154-5

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J. Y., Zhang, D., Engels, J. M. M., & Cronk, Q. (2012). Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany*, *99*(2), 320–329. https://doi.org/10.3732/ajb.1100570

Kerr, K. C. R., Lijtmaer, D. A., Barreira, A. S., Hebert, P. D. N., & Tubaro, P. L. (2009). Probing evolutionary patterns in neotropical birds through DNA barcodes. *PLoS ONE*, *4*(2), e4379. https://doi.org/10.1371/journal.pone.0004379

Langlois, V. S., Allison, M. J., Bergman, L. C., To, T. A., & Helbing, C. C. (2021). The need for robust qPCR-based eDNA detection assays in environmental monitoring and species inventories. *Environmental DNA*, *3*(3), 519–527. https://doi.org/10.1002/edn3.164

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656. https://doi.org/10.1073/pnas.1911714116

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2020). Reply to Locatelli et al.: Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group. *Proceedings of the National Academy of Sciences*, *117*(51), 32213–32214. https://doi.org/10.1073/pnas.2019903117

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: From gene to genome: Plant identification using DNA barcodes. *Biological Reviews*, *90*(1), 157–166. https://doi.org/10.1111/brv.12104

Lim, N. K. M., Tay, Y. C., Srivathsan, A., Tan, J. W. T., Kwik, J. T. B., Baloğlu, B., Meier, R., & Yeo, D. C. J. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science*, *3*(11), 160635. https://doi.org/10.1098/rsos.160635

Locatelli, N. S., McIntyre, P. B., Therkildsen, N. O., & Baetscher, D. S. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences*, *117*(51), 32211–32212. https://doi.org/10.1073/pnas.2007421117

Lynggaard, C., Bertelsen, M. F., Jensen, C. V., Johnson, M. S., Frøslev, T. G., Olsen, M. T., & Bohmann, K. (2022). Airborne environmental DNA for terrestrial vertebrate community monitoring. *Current Biology*, *32*(3), 701–707. https://doi.org/10.1016/j.cub.2021.12.014

Markle, D. F., & Tomelleri, J. R. (2016). A guide to freshwater fishes of Oregon. Oregon State University Press.

Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLOS ONE*, *14*(6), e0217084. https://doi.org/10.1371/journal.pone.0217084

Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, *9*(1), 214. https://doi.org/10.1186/1471-2164-9-214

Miao, W., Song, L., Ba, S., Zhang, L., Guan, G., Zhang, Z., & Ning, K. (2020). Protist 10,000 Genomes Project. *The Innovation*, *1*(3), 100058. https://doi.org/10.1016/j.xinn.2020.100058

Miya M., Sato Y., Fukunaga T., Sado T., Poulsen J. Y., Sato K., Minamoto T., Yamamoto S., Yamanaka H., Araki H., Kondoh M., & Iwasaki W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes:

Detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7), 150088. https://doi.org/10.1098/rsos.150088

Nakazawa, M. (2021). fmsb: Functions for medical statistics book with some demographic data. https://CRAN.R-project.org/package=fmsb

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., Green, R. E., & Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, *18*(5), 927–939. https://doi.org/10.1111/1755-0998.12895

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Olson, Z. H., Briggler, J. T., & Williams, R. N. (2012). An eDNA approach to detect eastern hellbenders (Cryptobranchus a. Alleganiensis) using samples of water. *Wildlife Research*, *39*(7), 629. https://doi.org/10.1071/WR12114

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830. https://doi.org/10.1111/1755-0998.12355

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*(2), 527–541. https://doi.org/10.1111/mec.13481

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338. https://doi.org/10.1111/mec.14478

Prendini, L., Hanner, R., & DeSalle, R. (2002). Obtaining, storing and archiving specimens and tissue samples for use in molecular studies. In R. DeSalle, G. Giribet, & W. Wheeler (Eds.), *Techniques in Molecular Systematics and Evolution* (pp. 176–248). Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-8125-8_11

Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rimet, F., Aylagas, E., Borja, Á., Bouchez, A., Canino, A., Chauvin, C., Chonova, T., Ciampor Jr, F., Costa, F. O., Ferrari, B. J. D., Gastineau, R., Goulon, C., Gugger, M., Holzmann, M., Jahn, R., Kahlert, M., Kusber, W. H., Laplace-Treyture, C., Leese, F., … Ekrem, T. (2021). Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. *Metabarcoding and Metagenomics*, *5*, e58056. https://doi.org/10.3897/mbmg.5.58056

Rowsey, D. M., & Egge, J. J. (2017). Morphometric analysis of two enigmatic sculpin species, *Cottus gulosus* and *Cottus perplexus* (Scorpaeniformes: Cottidae). *Northwestern Naturalist*, *98*(3), 190–202. https://doi.org/10.1898/NWN16-23.1

Salewski, V. (2003). Satellite species in lampreys: A worldwide trend for ecological speciation in sympatry?.. *Journal of Fish Biology*, *63*(2), 267–279. https://doi.org/10.1046/j.1095-8649.2003.00166.x

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2021). GGally: Extension to "ggplot2." https://CRAN.R-project.org/package=GGally

Schnell, I. B., Fraser, M., Willerslev, E., & Gilbert, M. T. P. (2010). Characterisation of insect and plant origins using DNA extracted from small volumes of bee honey. *Arthropod-Plant Interactions*, *4*(2), 107–116. https://doi.org/10.1007/s11829-010-9089-0

Shaffer, H. B., Toffelmier, E., Corbett-Detig, R. B., Escalona, M., Erickson, B., Fiedler, P., Gold, M., Harrigan, R. J., Hodges, S., Luckau, T. K., Miller, C., Oliveira, D. R., Shaffer, K. E., Shapiro, B., Sork, V. L., & Wang, I. J. (2022). Landscape genomics to enable conservation actions: The California Conservation Genomics Project. *Journal of Heredity*, *113*(6), 577–588. https://doi.org/10.1093/jhered/esac020

Sheffield, C. S., Hebert, P. D. N., Kevan, P. G., & Packer, L. (2009). DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. *Molecular Ecology Resources*, *9*, 196–207. https://doi.org/10.1111/j.1755-0998.2009.02645.x

Sipos, R., Székely, A., Révész, S., & Márialigeti, K. (2010). Addressing PCR biases in environmental microbiology studies. In S. P. Cummings (Ed.), *Bioremediation* (Vol. 599, pp. 37–58). Humana Press. https://doi.org/10.1007/978-1-60761-439-5_3

Smith, M. A., Poyarkov, N. A., & Hebert, P. D. N. (2008). DNA barcoding: CO1 DNA barcoding amphibians: take the chance, meet the challenge: DNA BARCODING. *Molecular Ecology Resources*, *8*(2), 235–246. https://doi.org/10.1111/j.1471-8286.2007.01964.x

Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H. M., & Schutten, M. (2010). The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5′ nuclease assay. *The Journal of Molecular Diagnostics*, *12*(1), 109–117. https://doi.org/10.2353/jmoldx.2010.090035

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., Yang, S., Moss, E. D., Wang, J., Yang, C., Bruce, C., Nevard, T., Potts, S. G., Zhou, X., & Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, *6*(9), 1034–1043. https://doi.org/10.1111/2041-210X.12416

Teeling, E. C., Vernes, S. C., Dávalos, L. M., Ray, D. A., Gilbert, M. T. P., Myers, E., & Bat1K Consortium. (2018). Bat biology, genomes, and the Bat1K Project: To generate chromosome-level genomes for all living bat species. *Annual Review of Animal Biosciences*, *6*(1), 23–46. https://doi.org/10.1146/annurev-animal-022516-022811

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, *47*(21), 10994–11006. https://doi.org/10.1093/nar/gkz841

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—New capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115–e115. https://doi.org/10.1093/nar/gks596

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942. https://doi.org/10.1111/mec.13428

Vladykov, V. D., & Kott, E. (1979). Satellite species among the holarctic lampreys (Petromyzonidae). *Canadian Journal of Zoology*, *57*(4), 860–867. https://doi.org/10.1139/z79-106

Waugh, J. (2007). DNA barcoding in animal species: Progress, potential and pitfalls. *BioEssays*, *29*(2), 188–197. https://doi.org/10.1002/bies.20529

Wu, J. H., Hong, P. Y., & Liu, W. T. (2009). Quantitative effects of position and type of single mismatch on single base primer extension. *Journal of Microbiological Methods*, *77*(3), 267–275. https://doi.org/10.1016/j.mimet.2009.03.001

Wydoski, R. S., & Whitney, R. R. (2003). Inland fishes of Washington (2nd ed., rev.expanded). American Fisheries Society in association with University of Washington Press.

Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., & Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, *7*(1), 40368. https://doi.org/10.1038/srep40368

Yang, C., Bohmann, K., Wang, X., Cai, W., Wales, N., Ding, Z., Gopalakrishnan, S., & Yu, D. W. (2021). Biodiversity Soup II: A bulk-sample metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*, *12*(7), 1252–1264. https://doi.org/10.1111/2041-210X.13602

Young, M. K., Smith, R., Pilgrim, K. L., Isaak, D. J., McKelvey, K. S., Parkes, S., Egge, J., & Schwartz, M. K. (2022). A molecular taxonomy of Cottus in western North America. *Western North American Naturalist*, *82*(2). https://doi.org/10.3398/064.082.0208

Young, M. R., Proctor, H. C., deWaard, J. R., & Hebert, P. D. N. (2019). DNA barcodes expose unexpected diversity in Canadian mites. *Molecular Ecology*, *28*(24), 5347–5359. https://doi.org/10.1111/mec.15292

Zemlak, T. S., Ward, R. D., Connell, A. D., Holmes, B. H., & Hebert, P. D. N. (2009). DNA barcoding reveals overlooked marine fishes. *Molecular Ecology Resources*, *9*, 237–242. https://doi.org/10.1111/j.1755-0998.2009.02649.x

Data Availability Statement

Appendix S3 contains all sequence data and details regarding the methodological processes used to derive the sequence data used in downstream analysis for this paper.

Table 1. Assembled Mitogenome Taxa Counts: OBGP specimens with assembled mitogenomes are grouped according to taxonomic designation with counts for each taxonomic level.

OBGP ID	Family		Genus		Species		Subspecies/Lineage	
OBGP-2019-237	Acipenseridae	3	*Acipenser*	3	*medirostris*	1		
OBGP-2018-244					*transmontanus*	2		
OBGP-2018-267								
OBGP-2019-023	Atherinopsidae	2	*Atherinops*	2	*affinis*	2		
OBGP-2019-230								
OBGP-2018-023	Catostomidae	61	*Catostomus*	48	*bondi*	3		
OBGP-2018-092								
OBGP-2019-038								
OBGP-2017-206					*columbianus*	1		
OBGP-2017-199					*macrocheilus*	2		
OBGP-2018-203								
OBGP-2017-020					*microps*	9		
OBGP-2017-021								
OBGP-2017-022								
OBGP-2017-090								
OBGP-2017-091								
OBGP-2017-092								
OBGP-2017-101								
OBGP-2017-103								
OBGP-2017-104								
OBGP-2017-007					*occidentalis*	8	*lacusanserinus*	5
OBGP-2017-008								
OBGP-2017-009								
OBGP-2017-011								
OBGP-2018-226								
OBGP-2017-093								
OBGP-2017-096								
OBGP-2017-100								
OBGP-2017-179					*rimiculus*	6	Jenny Creek	3
OBGP-2017-183								
OBGP-2017-184								
OBGP-2018-215							Klamath	2
OBGP-2019-156								
OBGP-2018-264							Rogue	1
OBGP-2017-232					*snyderi*	10		
OBGP-2017-233								
OBGP-2017-234								
OBGP-2017-235								
OBGP-2017-236								
OBGP-2017-237								
OBGP-2017-251								
OBGP-2017-252								
OBGP-2017-253								
OBGP-2017-254								
OBGP-2018-006					*tahoensis*	1		
OBGP-2019-142					*tsiltcoosensis*	6	Coos	2
OBGP-2019-148								
OBGP-2019-145							Coquille	1
OBGP-2017-216							Siuslaw	1
OBGP-2017-064							Umpqua	2
OBGP-2017-147								
OBGP-1993-001					*warnerensis*	2		
OBGP-2018-186								
OBGP-2017-302			*Chasmistes*	2	*brevirostris*	2		
OBGP-2017-304								
OBGP-2017-230			*Deltistes*	11	*luxatus*	11		
OBGP-2017-231								
OBGP-2017-249								
OBGP-2017-306								
OBGP-2017-309								
OBGP-2017-310								
OBGP-2017-311								
OBGP-2017-313								
OBGP-2017-314								
OBGP-2017-315								
OBGP-2017-316								
OBGP-2017-297	Centrarchidae	24	*Archoplites*	2	*interruptus*	2		
OBGP-2017-312								
OBGP-2017-178			*Lepomis*	10	*cyanellus*	1		
OBGP-2017-275					*gibbosus*	5		
OBGP-2017-320								
OBGP-2017-360								
OBGP-2018-042								
OBGP-2018-159								
OBGP-2018-172					*gulosus*	1		
OBGP-2017-277					*macrochirus*	3		
OBGP-2017-381								
OBGP-2018-174								
OBGP-2017-063			*Micropterus*	7	*dolomieu*	2		
OBGP-2017-151								
OBGP-2017-238					*salmoides*	5		
OBGP-2017-241								
OBGP-2017-276								
OBGP-2017-287								
OBGP-2019-057								
OBGP-2017-001			*Pomoxis*	5	*nigromaculatus*	5		
OBGP-2017-050								
OBGP-2017-308								
OBGP-2017-349								
OBGP-2018-177								
OBGP-2018-111	Clupeidae	2	*Alosa*	2	*sapidissima*	2		
OBGP-2018-179								
OBGP-2018-105	Cobitidae	2	*Misgurnus*	2	*anguillicaudatus*	2		
OBGP-2018-185								
OBGP-2017-326	Cottidae	46	*Cottus*	45	*aleuticus*	2		
OBGP-2017-350								
OBGP-2017-012					*asper*	4		
OBGP-2017-220								
OBGP-2017-269								
OBGP-2017-272								
OBGP-2019-178					*bairdii*	1		
OBGP-2017-148					*beldingii*	6	ssp.	3
OBGP-2018-320								
OBGP-2018-321								
OBGP-2017-203								
OBGP-2018-156								
OBGP-2018-287								
OBGP-2018-036					*bendirei*	1		
OBGP-2017-351					*confusus*	2		
OBGP-2018-012								
OBGP-2016-004					*gulosus*	5		
OBGP-2017-056								
OBGP-2017-084								
OBGP-2017-188								
OBGP-2017-346								
OBGP-2017-218					*klamathensis*	3		
OBGP-2017-246								
OBGP-2017-247								
OBGP-2018-127					*marginatus*	2		
OBGP-2018-138								
OBGP-2017-132					*perplexus*	10		
OBGP-2017-134								
OBGP-2017-138								
OBGP-2017-140								
OBGP-2017-192								
OBGP-2017-193								
OBGP-2017-270								
OBGP-2017-273								
OBGP-2017-285								
OBGP-2017-318								
OBGP-2019-138					*pitensis*	2		
OBGP-2019-150								
OBGP-2019-160					*polyporus*	1		
OBGP-2017-212					*princeps*	1		
OBGP-2017-141					*rhotheus*	3		
OBGP-2017-201								
OBGP-2017-288								
OBGP-2017-162					*tenuis*	2		
OBGP-2017-171								
OBGP-2019-025			*Enophrys*	1	*bison*	1		
OBGP-2017-200	Cyprinidae	79	*Acrocheilus*	2	*alutaceus*	2		
OBGP-2017-207								
OBGP-2017-239			*Carassius*	2	*auratus*	2		
OBGP-2017-240								
OBGP-2018-047			*Cyprinus*	2	*carpio*	2		
OBGP-2018-288								
OBGP-2017-170			*Gila*	3	*coerulea*	3		
OBGP-2017-244								
OBGP-2017-245								
OBGP-2018-221			*Hesperoleucus*	2	*symmetricus*	2		
OBGP-2018-223								
OBGP-2017-327			*Mylocheilus*	1	*caurinus*	1		
OBGP-2017-370			*Notemigonus*	1	*crysoleucas*	1		
OBGP-2017-032			*Oregonichthys*	5	*crameri*	2		
OBGP-2017-137								
OBGP-2016-002					*kalawatseti*	3		
OBGP-2017-099								
OBGP-2018-232								
OBGP-2017-176			*Pimephales*	3	*promelas*	3		
OBGP-2018-033								
OBGP-2018-216								
OBGP-2017-019			*Ptychocheilus*	5	*oregonensis*	2		
OBGP-2017-195								
OBGP-2017-135					*umpquae*	3	Siuslaw	1
OBGP-2017-154								
OBGP-2018-089								
OBGP-2017-054			*Rhinichthys*	26	*cataractae*	4	'Millicoma Dace'	1
OBGP-2017-014								
OBGP-2017-175								
OBGP-2018-242								
OBGP-2016-001					*evermanni*	2		
OBGP-2018-184								
OBGP-2017-330					*falcatus*	2		
OBGP-2018-100								
OBGP-2016-005					*osculus*	16	Black Lined	5
OBGP-2017-016								
OBGP-2017-017								
OBGP-2017-279								
OBGP-2017-290								
OBGP-2017-086							Closed Basin	2
OBGP-2018-019								
OBGP-2018-189							Foskett Spring	2
OBGP-2018-190								
OBGP-2017-166							Klamath	2
OBGP-2017-172								
OBGP-2018-057							Malheur Stream	2
OBGP-2018-061								
OBGP-2018-045							Stinking Lake Spring	1
OBGP-2017-158							Western	2
OBGP-2017-202								
OBGP-2018-069					*umatilla*	2		
OBGP-2018-122								
OBGP-2017-033			*Richardsonius*	14	*balteatus*	12	*balteatus*	5
OBGP-2017-197								
OBGP-2017-268								
OBGP-2017-359								
OBGP-2018-039								
OBGP-2018-005							*hydrophlox*	2
OBGP-2018-048								
OBGP-2016-006							Siuslaw	5
OBGP-2017-065								
OBGP-2017-278								
OBGP-2019-137								
OBGP-2019-149								
OBGP-2018-028					*egregius*	2		
OBGP-2018-094								
OBGP-2011-001			*Siphateles*	12	*alvordensis*	2		
OBGP-2019-212								
OBGP-2018-066					*bicolor*	8	*eurysoma*	1
OBGP-2017-177							Klamath	2
OBGP-2019-136								
OBGP-2018-007							*obesus*	2
OBGP-2018-026								
OBGP-2017-366							Silver Lake Basin	1
OBGP-2017-002							*thalassinus*	2
OBGP-2017-003								
OBGP-2009-001					*boraxobius*	2		
OBGP-2009-002								
OBGP-2019-223			*Tinca*	1	*tinca*	1		
OBGP-2017-055	Embiotocidae	3	*Cymatogaster*	1	*aggregata*	1		
OBGP-2019-013			*Embiotoca*	1	*lateralis*	1		
OBGP-2019-021			*Phanerodon*	1	*furcatus*	1		
OBGP-2019-224	Esocidae	1	*Esox*	1	*lucius*	1		
OBGP-2017-329	Fundulidae	2	*Fundulus*	2	*diaphanus*	2		
OBGP-2018-098								
OBGP-2017-136	Gasterosteidae	1	*Gasterosteus*	1	*aculeatus*	1		
OBGP-2017-185	Ictaluridae	10	*Ameiurus*	7	*natalis*	3		
OBGP-2017-305								
OBGP-2017-307								
OBGP-2017-018					*nebulosus*	4		
OBGP-2017-257								
OBGP-2018-046								
OBGP-2019-170								
OBGP-2018-292			*Ictalurus*	2	*punctatus*	2		
OBGP-2018-293								
OBGP-2019-194			*Noturus*	1	*gyrinus*	1		
OBGP-2019-032	Osmeridae	2	*Hypomesus*	1	*pretiosus*	1		
OBGP-2019-222			*Thaleichthys*	1	*pacificus*	1		
OBGP-2018-178	Oxudercidae	1	*Rhinogobius*	1	*brunneus*	1		
OBGP-2017-221	Percidae	4	*Perca*	3	*flavescens*	3		
OBGP-2017-242								
OBGP-2017-243								
OBGP-2018-268			*Sander*	1	*vitreus*	1		
OBGP-2017-383	Percopsidae	1	*Percopsis*	1	*transmontana*	1		
OBGP-2016-007	Petromyzontidae	13	*Entosphenus*	12	*lethophagus*	3		
OBGP-2017-024								
OBGP-2017-025								
OBGP-2017-027					*minimus*	1		
OBGP-2017-248					*similis*	3		
OBGP-2017-250								
OBGP-2019-143								
OBGP-2017-030					*tridentatus*	5		
OBGP-2017-325								
OBGP-2019-058								
OBGP-2019-167								
OBGP-2019-168								
OBGP-2019-010			*Lampetra*	1	*richardsoni*	1		
OBGP-2019-027	Pholidae	2	*Apodichthys*	1	*flavidus*	1		
OBGP-2019-029			*Pholis*	1	*ornata*	1		
OBGP-2019-026	Pleuronectidae	3	*Parophrys*	1	*vetulus*	1		
OBGP-2017-053			*Platichthys*	2	*stellatus*	2		
OBGP-2018-181								
OBGP-2017-342	Poeciliidae	2	*Gambusia*	2	*affinis*	2		
OBGP-2019-060								
OBGP-2018-065	Salmonidae	47	*Oncorhynchus*	27	*clarkii*	8	*henshawi*	3
OBGP-2018-068								
OBGP-2018-096								
OBGP-2017-149							*lewisi*	2
OBGP-2017-155								
OBGP-2017-332								
OBGP-2017-348								
OBGP-2018-240								
OBGP-2017-258					*keta*	2		
OBGP-2017-259								
OBGP-2017-271					*kisutch*	3		
OBGP-2018-038								
OBGP-2018-248								
OBGP-2017-015					*mykiss*	9	*gairdneri*	3
OBGP-2017-194								
OBGP-2017-198								
OBGP-2016-003							*irideus*	3
OBGP-2017-013								
OBGP-2017-061								
OBGP-2017-180							*newberrii*	3
OBGP-2017-255								
OBGP-2017-256								
OBGP-2019-190					*nerka*	2		
OBGP-2019-191								
OBGP-2017-052					*tshawytscha*	3		
OBGP-2017-356								
OBGP-2019-056								
OBGP-2017-062			*Prosopium*	2	*williamsoni*	2		
OBGP-2017-196								
OBGP-2017-167			*Salmo*	3	*trutta*	3		
OBGP-2017-227								
OBGP-2017-229								
OBGP-2011-002			*Salvelinus*	15	*confluentus*	6		
OBGP-2014-001								
OBGP-2017-223								
OBGP-2019-269								
OBGP-2019-272								
OBGP-2019-273								
OBGP-2017-168					*fontinalis*	6		
OBGP-2017-226								
OBGP-2017-228								
OBGP-2017-368								
OBGP-2018-020								
OBGP-2018-064								
OBGP-2019-211					*namaycush*	3		
OBGP-2019-227								
OBGP-2019-228								
OBGP-2019-030	Sebastidae	1	*Sebastes*	1	*caurinus*	1		
OBGP-2019-008	Umbridae	1	*Novumbra*	1	*hubbsi*	1		

Table 2. Taxonomically Diagnostic Nucleotides (TDN) within Families: For each of 7 families, maximum and mean TDNs in a 150 base window shifted at 20 base intervals along an intrafamily alignment of mitochondrial gene regions are listed here. TDN "spikes", where max(TDN) > 2 * mean(TDN) are in bold. Proportional relationship between mean within-family intraspecies and interspecies identity (id_prop) is also listed.

	Salmonidae			Cyprinidae			Catostomidae			Centrarchidae			Cottidae			Ictaluridae			Petromyzontidae		
	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop	max	mean	id_prop
rrnS	**21**	8.05	0.971	17	9.22	0.961	8	4.28	0.991	24	14.4	0.909	**18**	8.07	0.983	24	15.3	0.948	7	4.03	0.995
rrnL	**31**	7.73	0.970	**38**	16.0	0.926	**13**	5.48	0.985	**38**	17.7	0.904	**21**	8.82	0.976	30	15.2	0.944	**8**	3.71	0.997
nad1	30	23.5	0.882	28	19.0	0.849	29	16.2	0.950	42	30.8	0.827	26	18.6	0.938	48	33.0	0.868	24	13.9	0.988
nad2	47	35.0	0.871	29	18.2	0.826	23	13.8	0.940	51	36.1	0.795	35	24.3	0.927	44	33.8	0.861	30	18.7	0.984
cox1	29	17.6	0.908	**31**	14.4	0.892	18	9.01	0.973	38	26.1	0.853	24	17.5	0.957	41	29.5	0.883	19	10.8	0.991
cox2	32	18.8	0.932	19	13.4	0.906	13	9.54	0.977	36	21.8	0.869	25	16.7	0.964	27	17.4	0.911	20	14.4	0.987
cox3	27	20.3	0.912	21	13.5	0.890	17	12.3	0.970	35	25.8	0.855	24	14.9	0.956	39	27.9	0.896	18	12.7	0.989
nad4	35	25.7	0.885	33	19.7	0.846	24	17.4	0.954	50	31.3	0.809	28	17.6	0.938	44	34.4	0.860	23	15.1	0.987
nad5	43	23.3	0.889	39	22.0	0.832	22	13.6	0.958	**59**	28.5	0.812	35	18.4	0.940	51	34.0	0.862	23	15.3	0.985
cytb	33	21.3	0.894	24	15.7	0.864	18	13.2	0.952	51	31.7	0.826	22	16.9	0.952	41	31.9	0.873	18	12.5	0.989
dloop	**118**	38.8	0.792	**104**	57.0	0.791	**21**	9.93	0.948	**107**	31.0	0.834	**103**	34.7	0.919	**46**	22.0	0.759	**46**	15.6	0.982

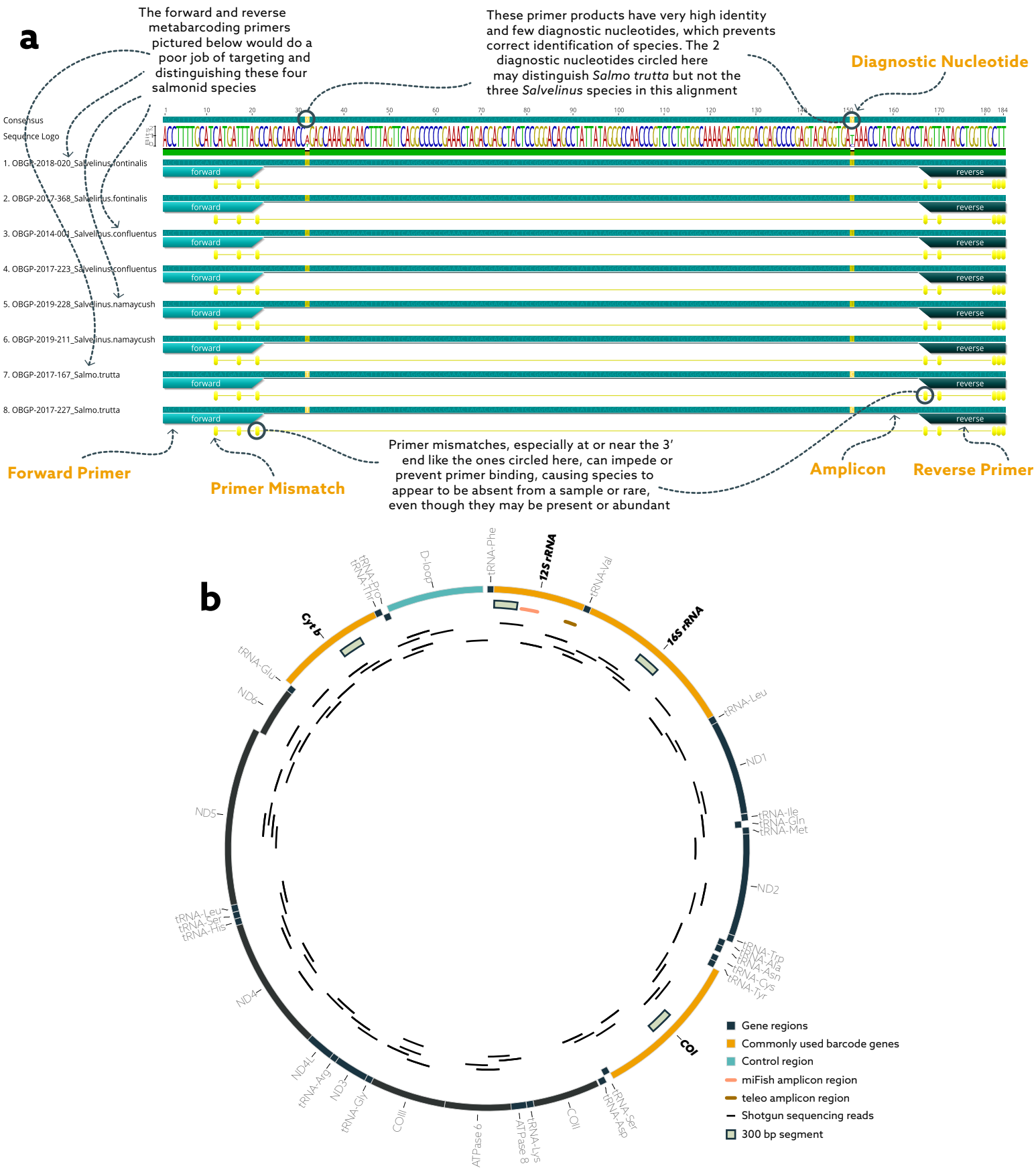
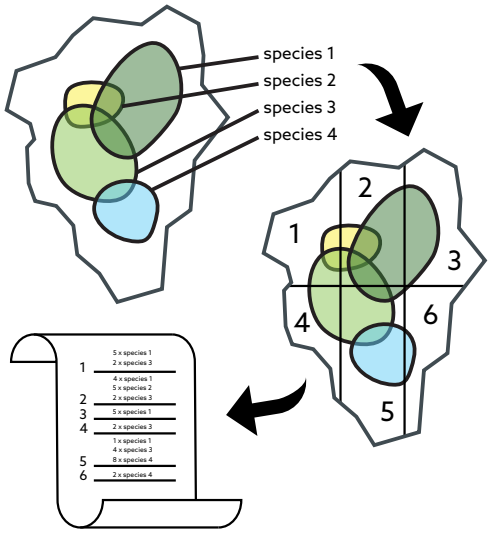


FIGURE 1
Primer Issues and Mapping Shotgun Reads. a. The Problems with Primers. A sequence alignment of PCR products produced *in silico* illustrates why amplicons produced by some primer pairs, like the ones depicted here, might fail to detect or identify species. b. Mapping Shotgun Reads. Samples that are shotgun-sequenced will produce reads from the entire mitogenome. This graphic depicts how, with whole mitogenomic information, it is possible to make use of all the data produced from shotgun sequencing. The length of barcodes is limited due to technological constraints and can only harness a small (≤ 300 bp) fragment of genetic information from the mitogenome while far more genetic information can be accessed by shotgun sequencing all the DNA that is present.

FIGURE 2

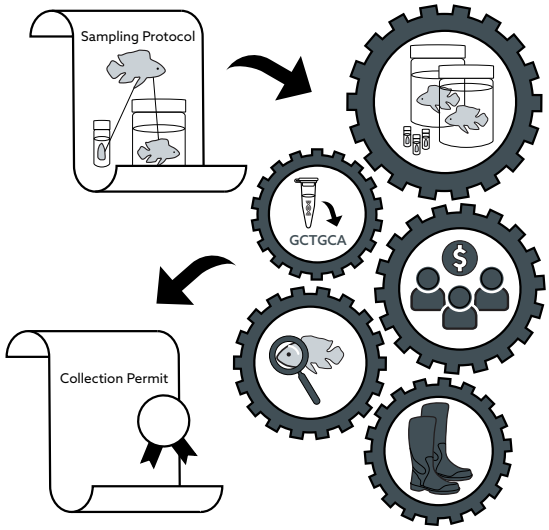
Creating a Reference Sequence Database for Monitoring with eDNA

A one-page blueprint designed to provide researchers with the basic steps to follow to create a regional database of full mitogenome sequences for target taxa.



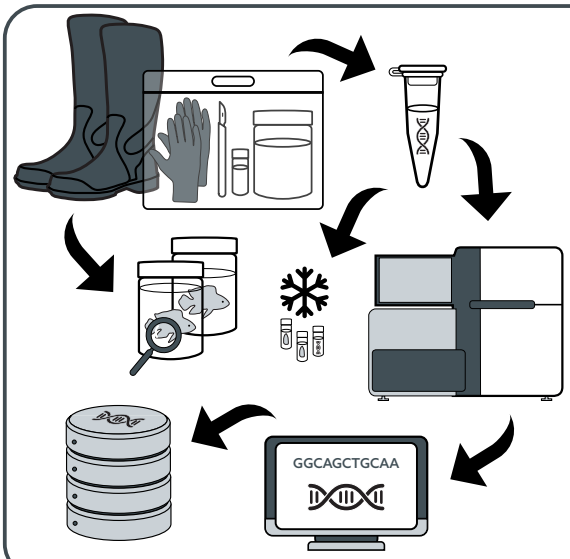
- Identify all target species present in your region of interest and determine their ranges.
- Divide your region into manageable subregions for sampling.
- Create a sampling plan to collect 3-10 individuals per species throughout its range.

- Within our study area, the state of Oregon, we identified 146 freshwater fish species and members of species complexes, both native and nonnative as resident target taxa.
- Because we were collecting freshwater species, it made sense to divided our study area into hydrologic units. Six parent regions—Coast, Lower Columbia, Willamette, Middle Columbia, Snake, and Closed Basins—were divided into 34 smaller basins.
- We set a goal of collecting 10 individuals of each species throughout its range.



- For your target taxa, research standard protocols for collecting and storing samples and, if taxonomic verification is needed, vouchers for downstream DNA analysis. Draft a sampling protocol and field datasheet for metadata.
- Gather your team, strategize collection, draft budget, seek funding. Identify where vouchers, if collected, samples and DNA extracts will be stored and taxonomically verified. Develop a wet-lab pipeline appropriate for your target taxa—adapted for HMW DNA if possible—locate a genetics lab and sequencing facilities.
- Apply for collection permits, if required.

- We developed a sampling protocol—included in SI—for the humane collection of fishes and stored tissues in 95% EtOH at -80°C to ensure proper preservation for DNA analysis. Our Field Notes Sheet is included in SI.
- Our core team included researchers from Oregon State University, the US Forest Service, and the Oregon Department of Fish and Wildlife (ODFW). We used the Cronn Lab for DNA extraction and library preparation, the Center for Qualitative Life Sciences for sequencing, and the Oregon State Ichthyology Collection (OSIC) to verify taxonomic assignments and store vouchers, tissues, and extracts.
- Our collection permits were managed by ODFW.



- Purchase supplies, assemble collection kits, and get boots on the ground to begin collection. Distribute samples to genetics lab and accession voucher specimens.
- Extract DNA, using HMW extraction protocols if possible. Accession samples—generally tissues or blood—and extracts and store at -80°C.
- Sequence samples once enough extracts are ready to fill a sequencing lane with sufficient coverage for your target taxa's mtDNA.
- Assemble mitogenome sequences using an appropriate bioinformatics pipeline.
- Store sequences on GenBank. Use a purpose-built client-server database, if desired.

- Details about our collection kit are included in SI. Tissues, extracts, and full-body vouchers are stored and taxonomically verified with OSIC.
- We extracted DNA from subsampled tissues using the Qiagen DNeasy Blood and Tissue Kit spin-column protocol for animal tissues. For more details, see Methods.
- We sequenced a mean of 67 samples per lane on an Illumina HiSeq 3000.
- SPAdes assembler successfully resolved most of our mitogenome sequences.
- Our data are available on GenBank and in SI, and obgpdb.org is reserved for a stand-alone client-server database.

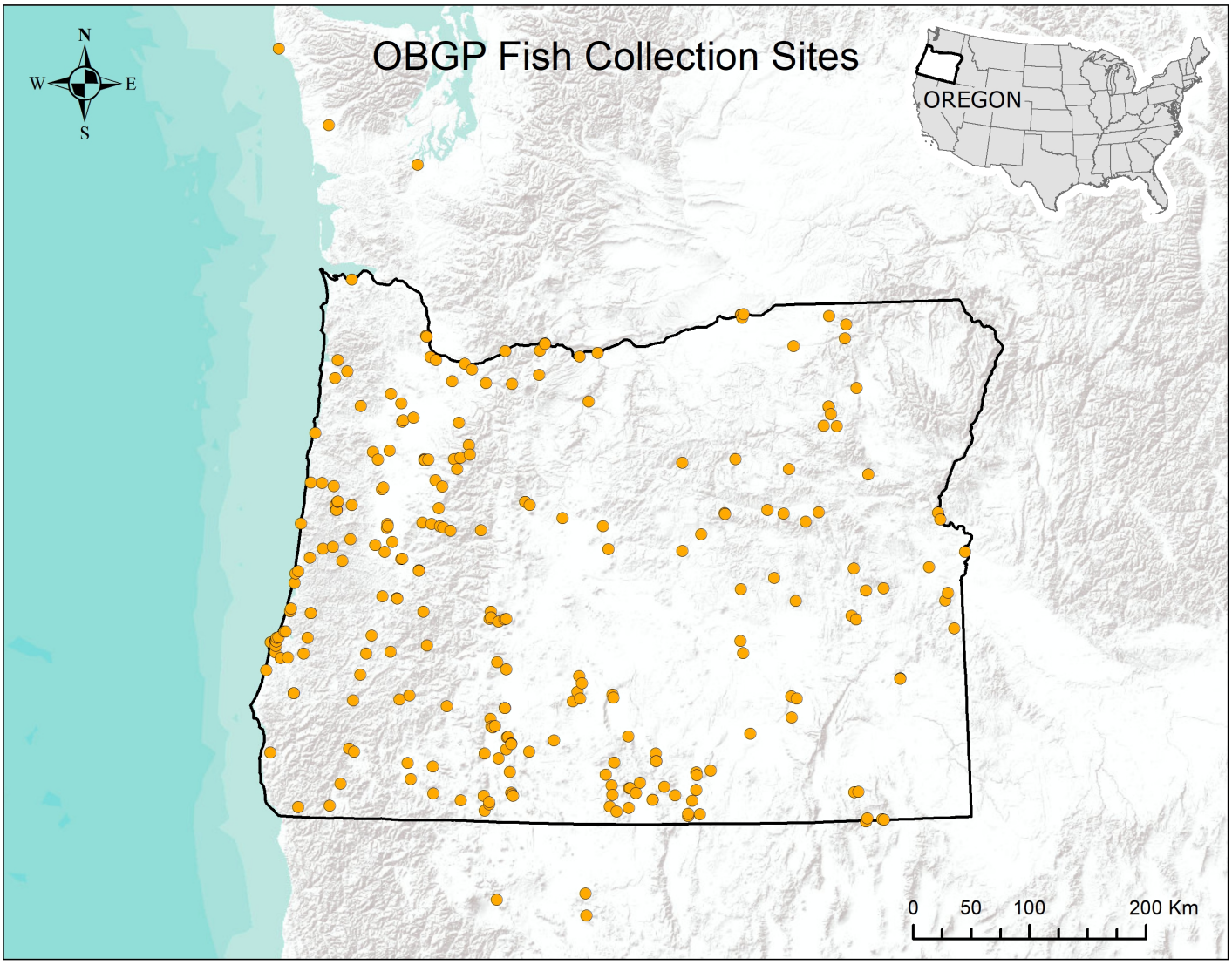


FIGURE 3
Map of Study Area and Sampling Sites. Each orange circle represents a single sampling location. An interactive map can be viewed here.

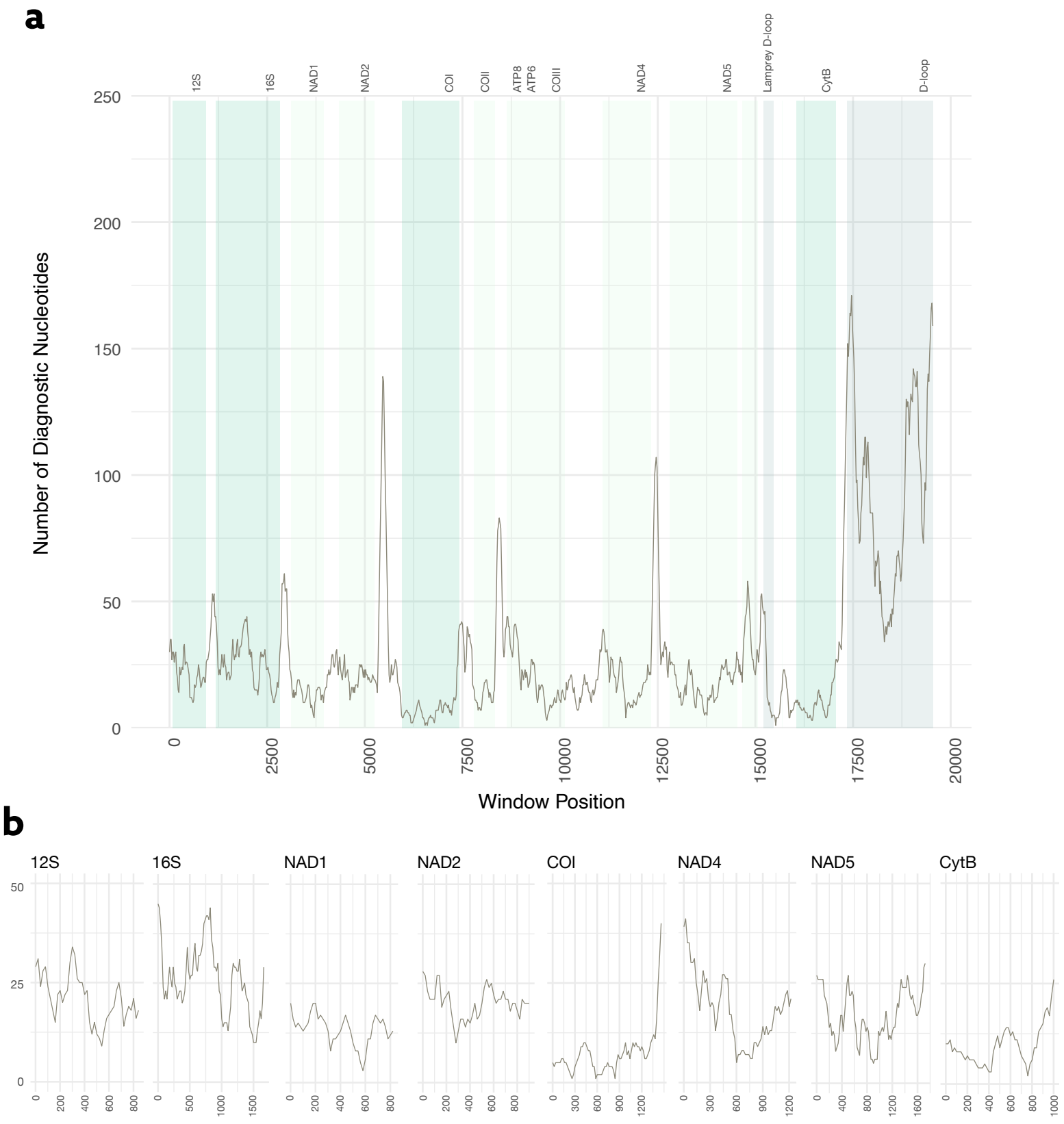


FIGURE 4

Sliding Window Analysis: A window 150 bases in length is placed at the beginning of an alignment of 313 mitogenome sequences and shifted right at 20 base intervals. At each window position, the number of taxonomically diagnostic nucleotides (TDN)—where a nucleotide is shared within a species but is either different or unaligned with other species—is counted. Areas with high concentrations of TDNs are likely to be superior for species identification in eDNA assays. a. Full Mitogenome: Gene regions and the D-loop are shaded in blue. Areas with spikes in TDNs are located across the entire mitogenome and the highest concentrations of TDNs are located in noncoding regions. b. Individual Genes: These plots zoom in on a subset of individual genes within the mitogenome to focus on the number of diagnostic nucleotides within 8 barcode regions. The 16S region has the highest taxonomically diagnostic nucleotides per 150-base window shifted at 20-base intervals. Mean TDN/$w_{150}i_{20}$ in each region: COI, 7.257; CytB, 9.451; NAD1, 13.381; NAD5, 17.092; NAD4, 18.226; NAD2, 20.065; 12S, 20.814; 16S, 25.726

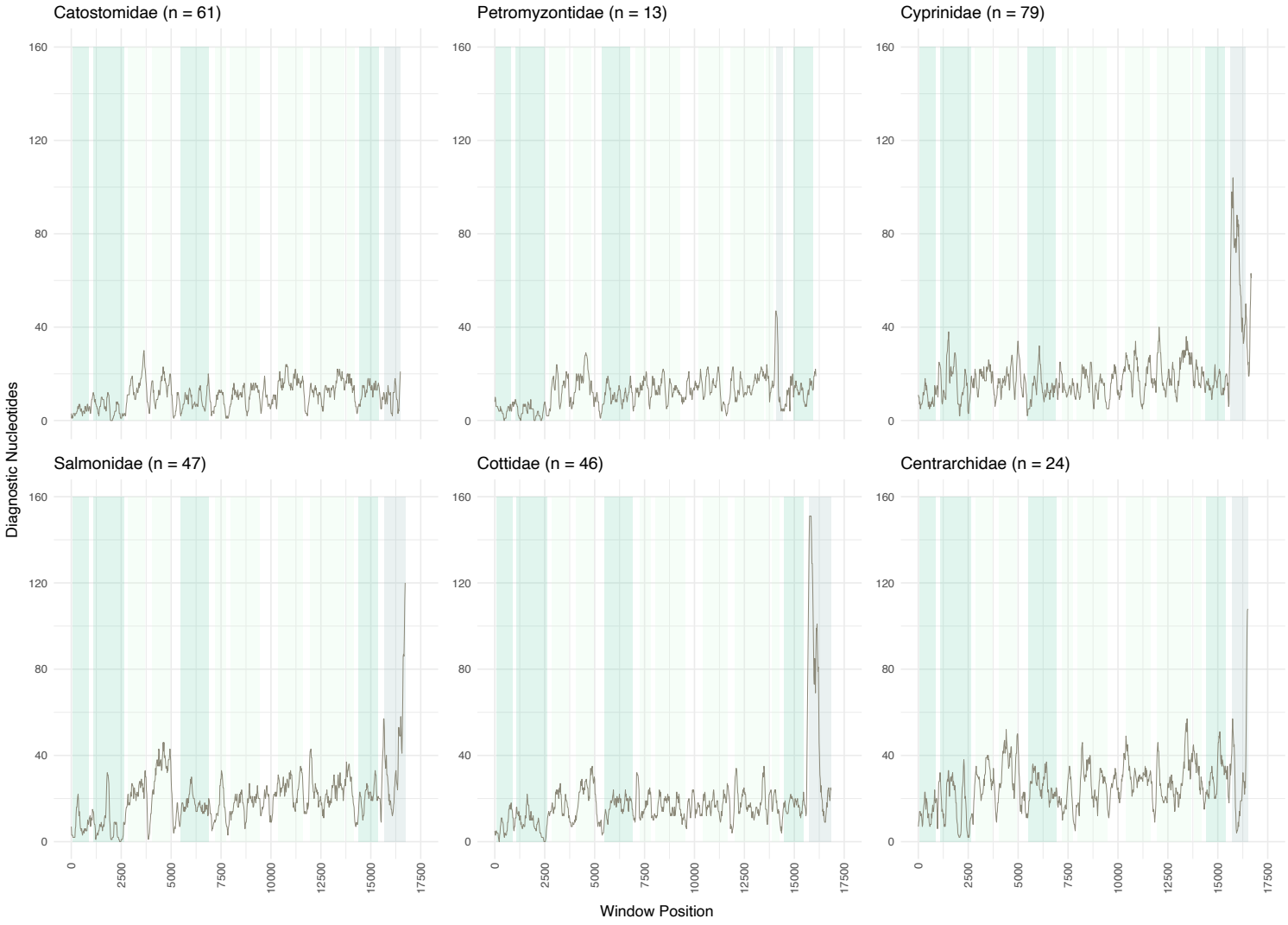


FIGURE 5
Sliding Window Analysis by Family, Full Mitogenome: Sequences from families with mitogenomic sequence information for >4 species and multiple specimens for each species are depicted here. Sequences from each family are separated into within-family groups and realigned. A window 150 bases in length is placed at the beginning of each within-family alignment and shifted right at 20 base intervals. At each window position, the number of taxonomically diagnostic nucleotides—where a base is shared within a species but is either different or unaligned with other species—is counted. This plot illustrates that different families have different levels of interspecies variability—expressed as a concentration of taxonomically diagnostic nucleotides—across the mitogenome. Interspecies variability appears to be low in Catostomidae and relatively high in Centrarchidae suggesting centrarchid species will be easier to identify than catostomids in eDNA assays. Commonly used barcode genes are highlighted in deep bluegreen, from left to right, 12S, 16S, COI, CytB. Note: Petromyzontidae mitogenome is structured with its control region upstream of the CytB gene. Means (TDN/$w_{150}i_{20}$): Catostomidae, 10.656; Petromyzontidae, 12.104; Cottidae, 18.871; Cyprinidae, 19.316; Salmonidae, 20.209; Centrarchidae, 26.082.

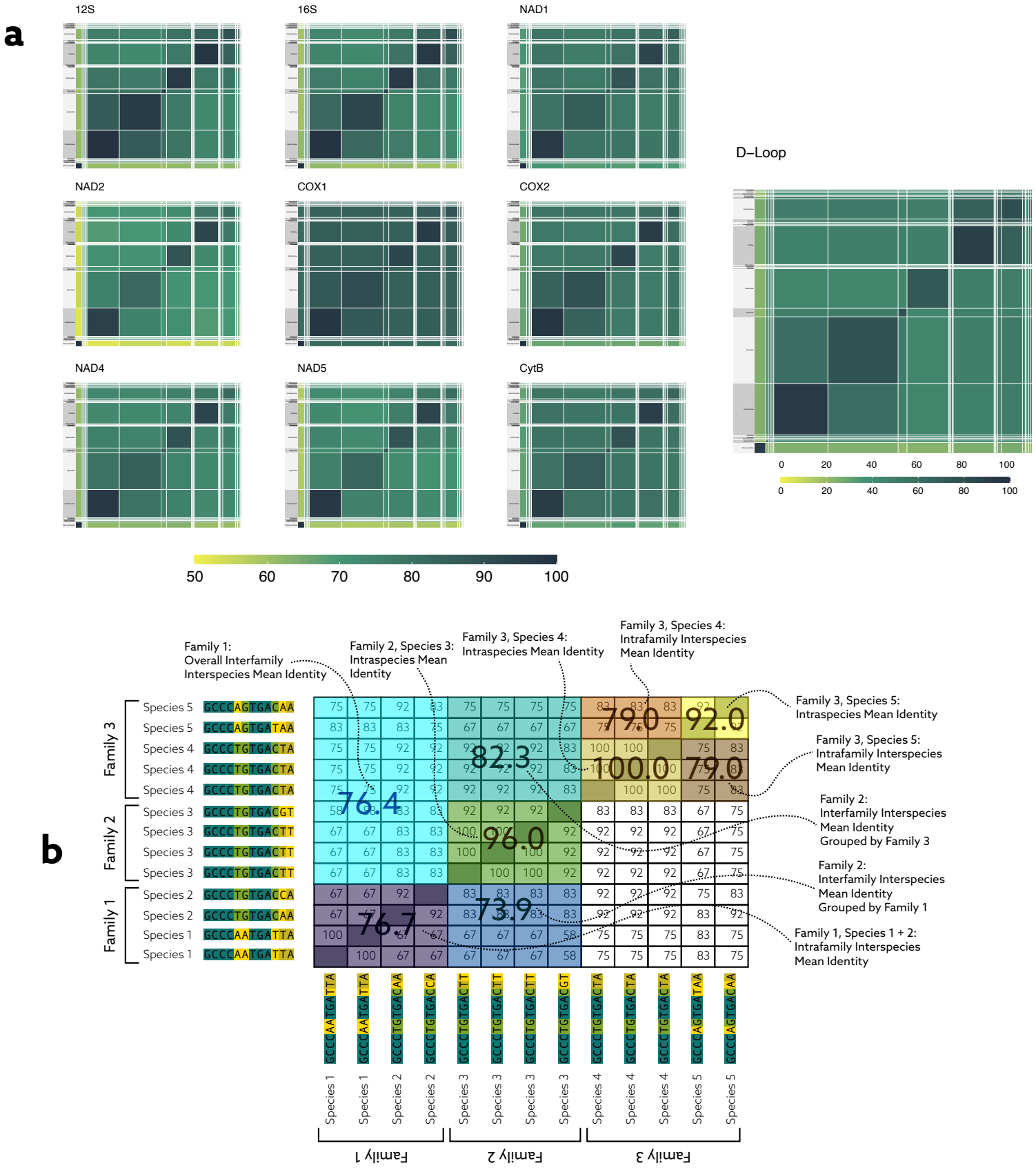


FIGURE 6

a. Percent Identity Heatmaps: Family level. This is a graphical representation of interfamily mitogenome percent identity. Yellow colors indicate greater dissimilarity while 100% identity is represented in dark purple; higher contrast therefore indicates greater distance in identity between families. Y- and X-axes are identical with each block on an axis representing one family. Raw numbers can be referenced in Appendix S5. A species-level heatmap is available in Appendix S4 Figure S9. b. Schematic View of an Identity Matrix: An alignment of sequences is compared in a pairwise fashion to determine the distance between one sequence and all other sequences in the alignment. The resulting matrix is symmetric along the diagonal with the central diagonal—where a sequence is compared to itself—remaining blank. Data can be grouped, and mean percent identities can be calculated as depicted here.

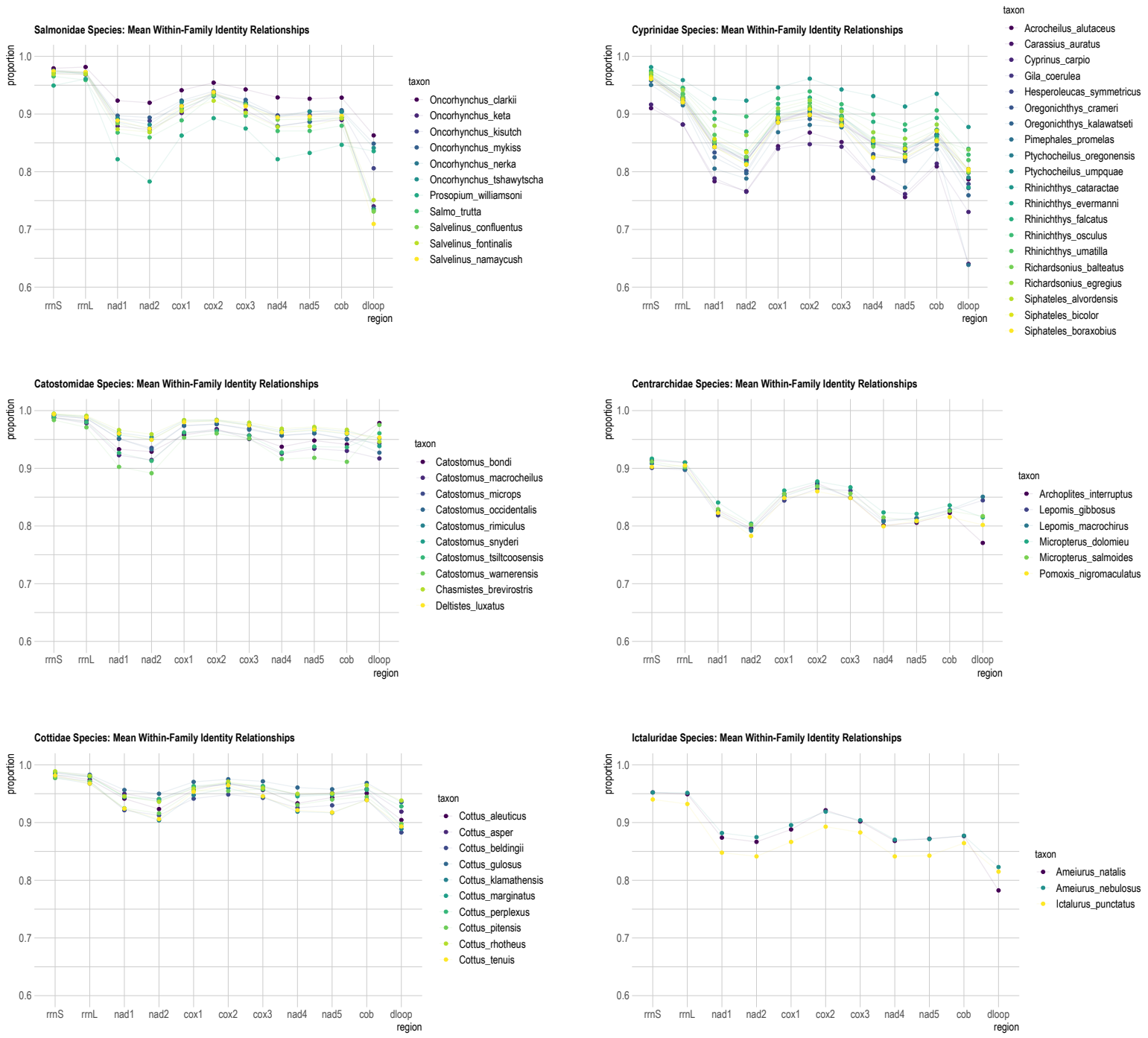


FIGURE 7

Within-Family Relationships in Mean Percent Identity: Parallel Coordinate Plots. These plots illustrate the relationship between intraspecies and interspecies genetic distance within the 5 plotted families at 11 regions (10 gene regions and the D-loop). At each of these 11 mitogenomic regions, mean intraspecies percent identity is calculated and then mean percent identity is calculated between that species and all other species within a given family. For each species at each region, the proportional relationship between these two means is plotted. A higher proportional value indicates higher genetic similarity between species within a family, so identifying species within these families may be more difficult using eDNA assays. Species within families exhibiting greater genetic distance here—Cyprinidae and Centrarchidae—should therefore be easier to identify using shorter regions of the mitogenome than species within families with higher genetic similarity—Cottidae and Catostomidae. See Table S4 for proportional identity figures.

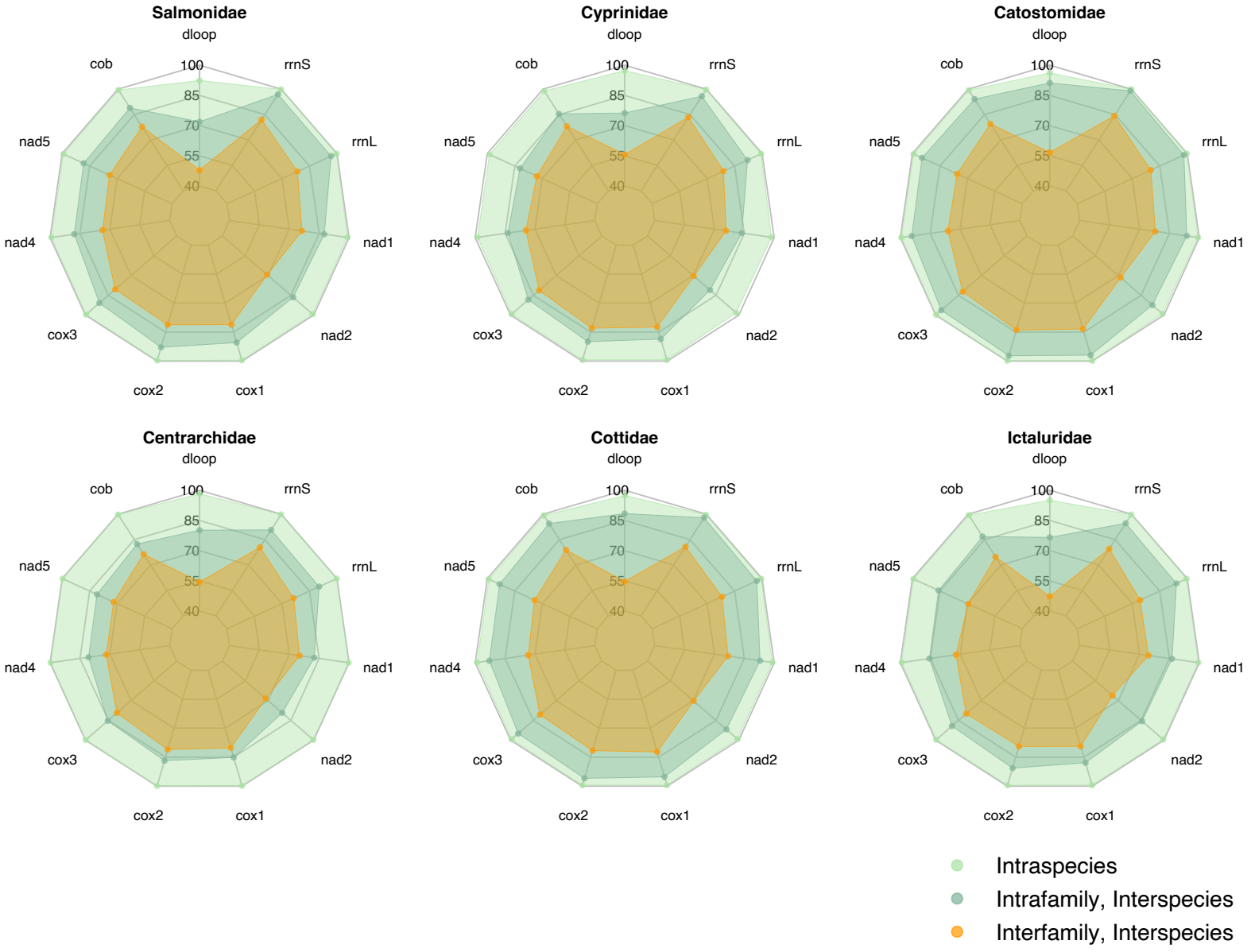


FIGURE 8
Percent Identity in a Subset of Families at Regions within the Mitogenome. Within each family, mean intraspecies percent identities are calculated and plotted for each gene and the D-loop (pale green), along with interspecies intrafamily percent identities (dark green), and interspecies interfamily percent identities (orange). Interfamily calculations are computed between species from all families, not just the families depicted here. Values on radar chart axes span from 40% identity at the innermost ring to 100% identity at the outermost ring. Genes are arranged in the order in which they occur in the circular mitogenome. Petromyzontidae is not included in plots due to highly skewed interspecies/interfamily identity. To identify species in eDNA assays, intraspecies identity should ideally be high and interspecies identity should be low. All families exhibit high intraspecies identity across all regions of the entire mitogenome with the lowest intraspecies identity found in the D-loop. Between families, there is sufficient interspecies variation to distinguish species from different families and identify sequences to the genus level. Within Cottidae and Catostomidae there is relatively high interspecies identity suggesting that species within these families may be difficult to identify. Within Salmonidae, there is high interspecies identity at the 12S (rrnS) and 16S (rrnL) gene regions suggesting these regions may not be ideal for identifying salmonid species.

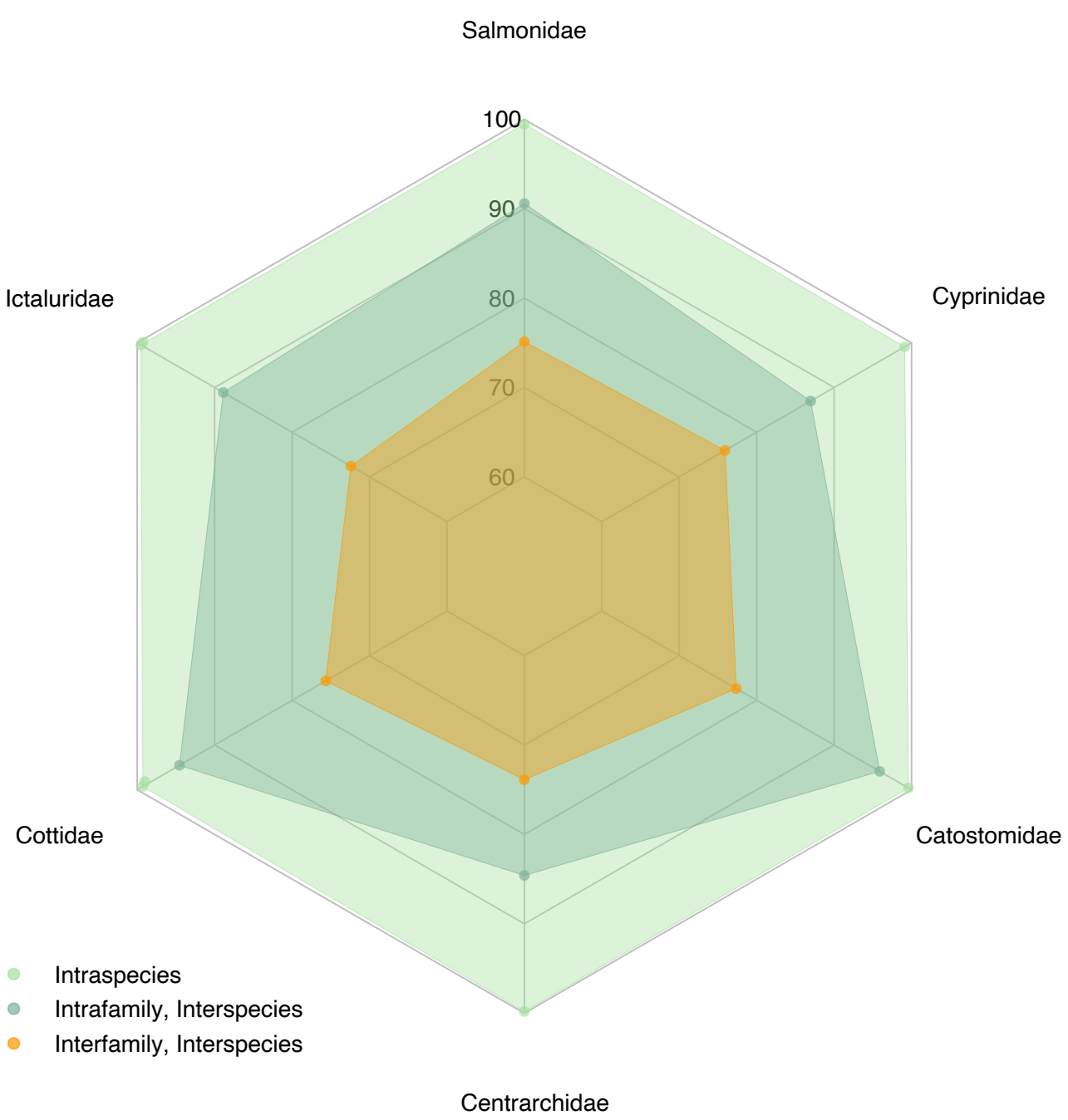


FIGURE 9
Whole Mitogenome Percent Identity in a Subset of Families. Within each family, mean intraspecies percent identities are calculated and plotted for the whole mitogenome (pale green), along with interspecies intrafamily percent identities (dark green), and interspecies interfamily percent identities (orange). Interfamily calculations are computed between species from all families, not just the families depicted here. Values on radar chart axes span from 60% identity at the innermost ring to 100% identity at the outermost ring. Petromyzontidae is not included in plots due to highly skewed interspecies/interfamily identity. To identify species in mitochondrial metagenomic assays, intraspecies identity should ideally be high and interspecies identity should be low. All families exhibit high intraspecies identity across the entire mitogenome. All families exhibit very low intrafamily interspecies identity and relatively low within-family interspecies identity suggesting the entire mitogenome should successfully identify species from all families.

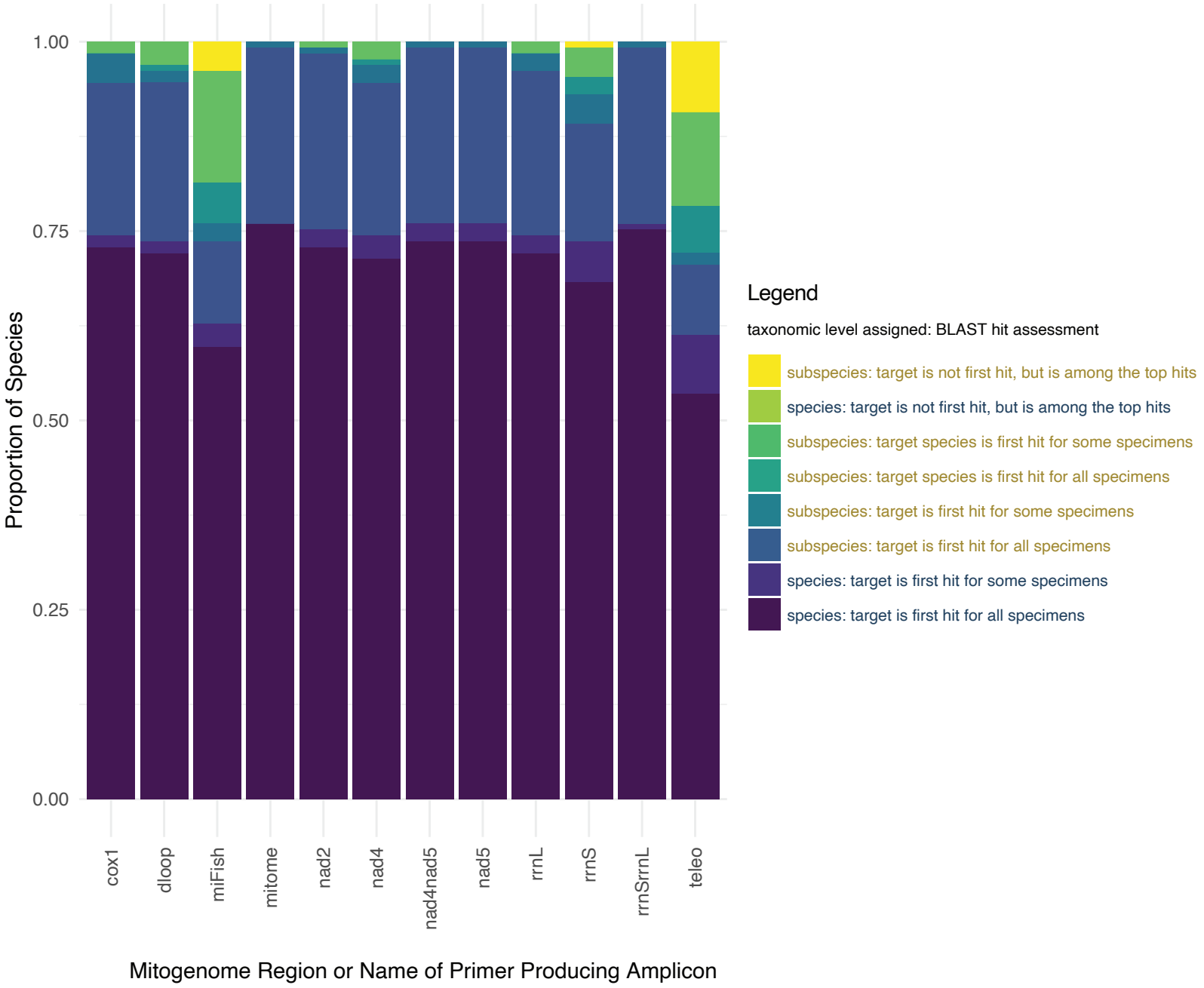


FIGURE 10
Mitogenome Regions and PCR Amplicons Queried Against Local OBGP Database: Analysis of All Highest E-Value BLAST Hits: Regions of the mitogenome were extracted from the alignment of 313 OBGP sequences and all sequence regions were blasted against a local database created from the same alignment. Single genes queried were CO1, NAD2, NAD4, NAD5, 16S (rrnL), and 12S (rrnS) regions. The entire region spanning from the beginning of the NAD4 to the end of the NAD5 (nad4nad5) genes and from the beginning of the 12S to the end of the16S (rrnSrrnL) genes were also queried along with the entire mitogenome (mitome) as well as the amplicons produced by both miFish and Teleo primer pairs. All hits with the highest E-Value measured in BLAST were analyzed to determine the proportion of species having the target species or subspecies as the first hit. Species or subspecies with multiple specimens represented in the alignment were evaluated to see if the target species or subspecies was the first hit for some or all representatives of that species. For the full mitogenome, the target species was the first hit for all specimens assigned to the species level and the target subspecies was the first hit for most subspecies that had specimens assigned to the subspecies level. For concatenated regions and the NAD5 gene, the results were similar to the full mitogenome, although the correct species was not always the first hit for all representatives of a species. The remaining regions were less likely to have the target as the first hit with miFish and Teleo amplicons having fewer first target hits than other, longer regions.